Statistical Methods for Studying Two Linked Disease Genes

by

Joanna Monika Biernacka

A thesis submitted in conformity with the requirements

for the degree of Doctor of Philosophy,

Graduate Department of Public Health Sciences,

University of Toronto

Abstract

Title:	Statistical methods for studying two linked disease genes
Degree:	Doctor of Philosophy
Year of convocation:	2005
Name:	Joanna Monika Biernacka
Department:	Graduate Department of Public Health Sciences
University:	University of Toronto

Assuming there is exactly one disease gene in a chromosomal region, a generalized estimating equations (GEE) approach can be used to estimate the location of the gene (Liang et al., 2001, Human Heredity 51: 64-78) using marker identical-by-descent (IBD) sharing data at multiple markers in a sample of affected sib pairs (ASPs). For diseases with complex genetic etiology, more than one susceptibility gene may exist in one chromosomal region. In such situations, linkage methods designed to detect a single locus may not successfully localize either of these two genes. We derived an expression for expected allele sharing in affected sib pairs at each point across a chromosomal segment containing two susceptibility genes, and proposed a GEE approach for localizing both disease genes simultaneously. We developed an algorithm that uses marker IBD sharing for a sample of ASPs to estimate the locations of the two genes and the expected ASP IBD sharing at these two loci. We also proposed methods to evaluate the evidence for two linked disease loci, in this GEE estimation framework, based on approximate quasi-likelihood ratio and generalized Wald and score test statistics. We evaluated the proposed estimation and testing methods by simulation, and found that the proposed estimation method can improve disease gene localization and aid in resolving large peaks when two disease genes are present in one chromosomal region. The performance of the estimation method for localizing two linked disease genes, and the power to detect the presence of two linked genes, improve with increased excess allele sharing at the disease gene loci, increased distance between the disease genes, and increased number of affected sib pairs. We applied the described methods to data from a genome scan for type 1 diabetes (Mein et al., 1998, Nature Genetics 19: 297-300) and obtained estimates of two putative disease gene locations on chromosome 6, approximately 20 cM apart.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Shelley Bull, for her constant support and guidance, and for all the opportunities she presented to me during my PhD studies. I would like to thank my committee members, Dr. Lei Sun and Dr. James Stafford, for their assistance and valuable comments throughout the course of this research. I would also like to acknowledge my examiners, Dr. Andrew Paterson and Dr. Kung-Yee Liang, as well as Dr. Paul Corey, and thank them for their thought-provoking questions and helpful remarks.

I am thankful to my colleagues and friends at the University of Toronto, for many helpful discussions and for making my years in the PhD program as much fun as educational. I would also like to thank my family, friends, and Steve for their encouragement, support, and absolute confidence in me.

I gratefully acknowledge the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory for allowing me to use their diabetes genome scan data (data version 2.0).

This research was partially supported by a Natural Sciences and Engineering Research Council of Canada graduate award, an Ontario Graduate Scholarship in Science and Technology, and an Ontario Student Opportunity Trust Fund award from the Samuel Lunenfeld Research Institute.

Table of Contents

Page

List of Tables

List of Figures

Page

List of Appendices

Chapter 1

Introduction

The aim of genetic linkage studies is the identification of genes which influence susceptibility to specific diseases. Various human traits and rare diseases, controlled by single genes and having simple inheritance patterns have been successfully studied for some time. Recently, however, research in human genetics has focused on investigating the more complex, but also more prevalent, oligogenic disorders and multifactorial traits, which may be controlled by both genetic and environmental factors, as well as their interactions. Current techniques in molecular genetics and DNA technology allow scientists to collect large amounts of genetic data, which makes genome-wide searches for multiple disease susceptibility genes more feasible. The data being gathered should be analyzed in the most efficient, effective, and accurate way possible. Application of appropriate statistical methods will further the development of knowledge in genetic research by allowing more efficient identification of genes responsible for various disorders with complex genetic bases.

1.1 Genetic Terminology and Basic Genetic Concepts

Before explaining some of the statistical methods used in human genetic research, we begin by providing definitions required for understanding genetic concepts in this thesis, as well as a brief description of the human genome, mechanism of genetic transmission from parents to offspring, and linkage analysis. Further details regarding these concepts can be found in introductory statistical genetic texts, such as Sham (1998), or review papers such as Olson et al. (1999).

The human genome consists of 23 pairs of chromosomes (22 pairs of autosomes and one pair of sex chromosome). Each *chromosome* is composed of two long strands of DNA (deoxyribonucleic acid) molecules normally bound to each other lengthwise and twisted in a double helix structure. A specific location along a given chromosome, or a specific position in the genome, is called a *locus*. The term *gene* is sometimes defined as a segment of DNA within a chromosome that specifies the amino acid sequence, and therefore the structure and function, of a single subunit of a protein. However, gene is also defined more broadly as any segment of DNA; thus the term gene is often used interchangeably with the term locus. The presence of different DNA sequences at the same locus in a population is known as *genetic polymorphism*. The alternative DNA sequences at a locus, or different “versions” of the same gene, are known as *alleles*.

Two chromosomes that belong to one of the pairs of chromosomes are said to be *homologous*. Thus, each of the 22 pairs of autosomes is made up of two homologous chromosomes. Consequently, humans have two copies of every gene found on the autosomal chromosomes, that is, two alleles of each autosomal gene. The two alleles that an individual

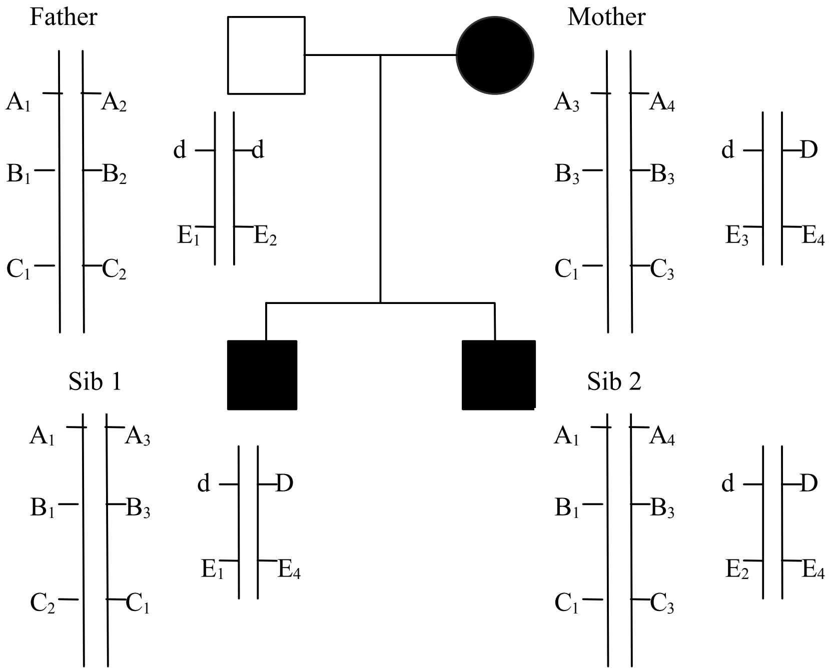


Figure 1.1: An example of transmission of alleles from parents to offspring. For simplicity only two chromosomes, with a total of five genetic loci, are shown.

possesses at a locus are referred to as their *genotype* at that locus. When the two alleles of an individual at some locus are identical, the genotype (or individual) is said to be *homozygous* for that gene, while if the two alleles for the same gene are different, that genotype is called *heterozygous*. For example, in Figure 1.1, the father is heterozygous at gene B, while the mother is homozygous at this gene. A *haplotype* is the pattern of alleles for a single chromosome involving more than one locus. For example, in Figure 1.1, the father has haplotypes $A_1B_1C_1$ and $A_2B_2C_2$ for genes A, B, and C on chromosome 1.

There are various ways that we can describe the effect of a gene on an observable trait, or *phenotype*. Let A_1 and A_2 be two different alleles at the same locus that influences a particular phenotype. If individuals with genotype A_1A_1 are phenotypically identical to individuals with genotype A_1A_2, but different from individuals with genotype A_2A_2, then allele A_1 is *dominant* to allele A_2, and A_2 is *recessive* to allele A_1. Alleles A_1 and A_2 are said to be *codominant* if the phenotype of individuals with genotype A_1A_2 is different from the phenotypes of both A_1A_1 and A_2A_2 individuals. The methods developed in this thesis are applicable to binary traits, such as disease status. The effect of a gene on a binary trait can be described by a *penetrance vector*. The *penetrance* of a given genotype is defined as the probability of a particular phenotype given the genotype value. In this thesis, the particular phenotype of interest will be disease status, and thus penetrance will mean probability of being affected, given the genotype at a disease susceptibility gene. For example, if a disease is caused by a mutation at a single gene, such that having two copies of the mutant allele (D) leads to disease, then the penetrances for the three genotypes d/d, d/D, D/D would be 0, 0, and 1, respectively. These three penetrances can be written as a penetrance vector (0,0,1). More generally, penetrance vectors will contain values between 0 and 1. For example, the penetrance vector (0.01,0.10,0.20) would indicate that even if a person does not have any copies of allele “D”, they may still get the disease with probability 0.01, for example because of environmental exposures or effects of other genes. If a person has one copy of allele “D” their probability of developing the disease increases to 0.1, while two copies of this allele increase this probability to 0.2. When two genes contribute to disease susceptibility, the disease-generating model can be described by a *penetrance matrix*; an example is shown in Figure 1.2. In this example, loci A and B interact to affect the probability of developing the disease. In other words, the phenotypic effect of the genotype at one of the genes depends on the genotype at the other gene. Generally, these types of interactions between two genes affecting one phenotype can be referred to as *epistasis*, and these genetic models are known as epistatic models. However, there is no clear consensus in the literature regarding the exact

		Genotype at disease gene A		
		A_1A_1	A_1A_2	A_2A_2
Genotype at disease gene B	B_1B_1	0.001	0.001	0.001
	B_1B_2	0.001	0.001	0.001
	B_2B_2	0.001	0.001	0.100

$\text{Pr(Disease|two-locus genotype} = A_2A_2B_2B_2)$

Figure 1.2: An example of a penetrance matrix for a two-locus epistatic model.

definition of the term "epistasis". See Cordell (2002) for a good discussion of this topic. Genetic models in which disease genes act independently to influence the disease, for example if the disease is caused by different genes in different populations, are referred to as *heterogeneity* models.

Analysis of genetic family data relies on an understanding of the mechanism of heredity. A brief introduction to this topic follows. An offspring receives one set of chromosomes from each parent, consisting of 22 autosomes and one sex chromosome. Mendel's first law, or the law of segregation, states that during reproduction, an offspring receives with equal probability one of the two alleles from the genotype of each parent. Alleles at two loci that are found on different chromosomes are transmitted independently of one another. This law of independent assortment (Mendel's second law) does not apply to some loci found on the same chromosome, because alleles on the same chromosome are more likely to be transmitted together. Typically, however, grand-parental chromosomes found in the parent are not transmitted to the offspring intact. Rather, during the production of gametes, segments of grand-parental homologous chromosomes are reshuffled during a process known as *crossing-over* resulting in recombination. Gametes are produced through a

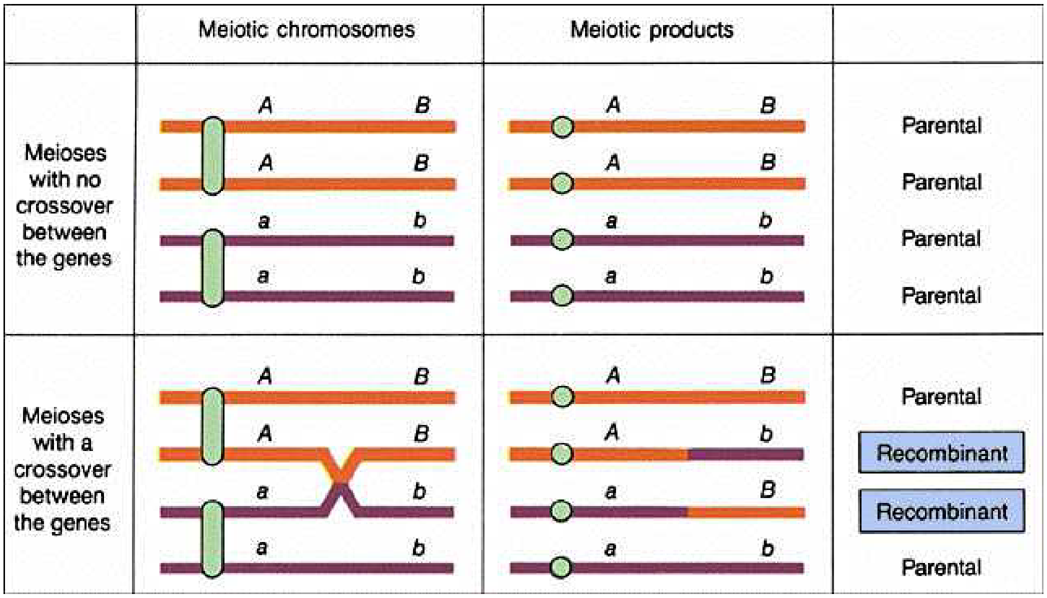

Meiotic chromosomes
Meiotic products
Meioses with no crossover between the genes
Meioses with a crossover between the genes
A
B
a
b
Parental
Recombinant

Figure 1.3: Recombination. Note that during meiosis (cell division that results in gamete production) each chromosome is duplicated producing two sister chromatids. Cross-overs between non-sister chromatid strands lead to recombinant chromosomes being transmitted to the offspring. (Griffiths et al. 1993, pg 123)

cell division process known as meiosis. At the beginning of meiosis each chromosome is duplicated creating a pair of sister chromatids. Cross-overs may then occur between two chromatids from the same pair of chromosomes as illustrated in Figure 1.3. The resulting exchange of genetic material between homologous chromosomes that produces a chromosome with alternating segments of paternally and maternally derived DNA is known as *recombination*.

Consider two loci: A and B. Assume that an individual inherits the haplotype A_mB_m from the mother, and A_fB_f from the father. Thus, the individual has genotype $A_fA_mB_fB_m$. The haplotype of a gamete produced by this individual will in general contain a mixture of alleles from that person's maternal and paternal haplotypes, for example A_mB_f. When the haplotype of the gamete is the same as one of the parental haplotypes (i.e. A_mB_m or A_fB_f), it is said to be a parental-type, or non-recombinant with respect to these two loci. On the other hand, if it contains one maternal allele and one paternal allele (i.e. A_mB_f or A_fB_m) then it is said to be non-parental or recombinant. The probability of a recombinant gamete is called the *recombination fraction*, commonly denoted by θ. Because of the independent assortment of chromosomes during gamete formation, θ is ½ for two loci found on different chromosomes. When two loci are on the same chromosome, separation of two maternal or two paternal alleles happens only if a crossover (or more precisely an odd number of crossovers) occurs between the two loci. The closer the two loci are to one another, the lower the probability of a crossover, and hence the lower the probability of a recombinant gamete, leading to a lower recombination fraction between the two loci.

The expected number of crossovers occurring between two loci on a single chromatid during meiosis is the *genetic map distance* between the two loci, measured in units of Morgans. The centiMorgan (cM), 1/100th of a Morgan, is the more commonly used map distance unit. A *map function* is a mathematical relationship that converts map distance (m measured in Morgans) to a recombination fraction (θ). The *Haldane function* is based on the assumption that crossovers occur at random independently of one another. The resulting function is $\theta = (1\text{-exp}(\text{-}2m))/2$. The Haldane map function does not appear to be accurate at small distances, because of a phenomenon known as *interference*, whereby the probability of two crossovers occurring in close proximity is lower than would be predicted by the Haldane map function. However, at larger distances interference becomes negligible and the Haldane map function is more accurate (Sham 1998, pg 55).

The process of recombination is of central importance to the statistical method of mapping genes known as linkage analysis. When the recombination fraction between two loci is lower than ½, the two loci are said to be *linked*. Linkage leads to co-segregation of alleles at two loci on the same chromosome. This phenomenon of linkage makes it possible to infer relative positions of two or more loci, by examining the patterns of allele co-transmission from parent to offspring at the loci. In the context of gene mapping, the goal of *linkage analysis* is to infer the position of a gene that contributes to a specific trait, by studying the co-segregation of the trait with alleles at polymorphic loci with known chromosomal locations, known as *genetic markers*.

1.2 Linkage Analysis

In the absence of good candidate genes, multipoint linkage analysis is the most commonly used technique as a first step in finding disease susceptibility genes. Usually a genome scan approach is taken, and all chromosomes are searched for regions containing disease susceptibility genes. Linkage analysis methods can be classified into two broad categories. The first category, commonly known as *parametric* or *model-based* linkage analysis, requires the specification of an underlying genetic model. With the assumed genetic model and the pedigree structures, the likelihood of the observed data can be constructed, as a function of the recombination fraction between a marker and the putative disease gene. The null hypothesis of no linkage ($\theta = 0.5$) can then be tested. The test statistic usually used in model-based linkage analysis is the LOD score (base 10 logarithm of the likelihood ratio). The second category of linkage analysis methods, *allele-sharing* methods (also known as *non-parametric* or *model-free* methods), do not require the assumption of a specific genetic model. Because these methods are not as dependent on assumptions about the underlying genetic model (such as penetrances), they are preferred by many researchers at early stages in the study of complex traits (e.g. see Ott 1996; Elston 2000). Since the methods developed in this thesis are based on allele-sharing, only this type of linkage analysis will be described.

1.2.1 Traditional model-free or allele-sharing-based linkage analysis

Markers having some alleles shared by the affected family members and other alleles shared by the unaffected members are likely to be located near disease susceptibility genes. This association between the sharing of disease status and the sharing of marker alleles by sets of relatives is the basis of model-free/allele-sharing linkage analysis methods. Allele-sharing linkage methods for binary traits usually use data from pairs (or sets) of affected relatives, most commonly affected sib-pairs. Because the methods described in this thesis are developed for affected sib-pairs, the discussion that follows will focus on this design.

Two alleles are said to be *identical by descent* (IBD) if they are not only of the same allelic form, but also are descended from the same ancestral allele. For example, the two A_1 alleles of the two siblings in Figure 1.1 are identical by descent, whereas their two C_1 alleles are identical by state, but not identical by descent. In some cases, IBD sharing is known based on the observed marker genotypes. For example, in Figure 1.1, it can be unambiguously determined that the two siblings share 1 allele IBD at gene A, and 0 alleles IBD at gene C. In such cases the marker data for the sib-pair is said to be *fully informative*. More generally, IBD sharing at a genetic marker cannot be unequivocally determined from genotype data, but rather, is determined probabilistically. In such situations, when the marker for a sib pair is not fully informative, the probability that the sib-pair shares *i* alleles IBD can be estimated conditional on the available marker data. In the example shown in Figure 1.1, because the two siblings share 1 allele IBD at gene A, and gene B is linked to gene A, it is more likely that the two siblings share 1 allele (rather than 2) IBD at gene B. In methods known as *multipoint linkage analysis*, multiple markers are used to probabilistically determine the number of alleles shared IBD at genetic markers, increasing the overall marker informativity. The software GENEHUNTER is one of many programs that can carry out such multipoint calculations to estimate IBD sharing probabilities for pairs of relatives. GENEHUNTER uses genotype information at multiple marker loci from all available family members to derive the probability distribution of all possible IBD configurations at a given locus, via the Lander-Green algorithm (Lander and Green 1987) based on a hidden Markov formulation of the pattern of inheritance at several loci.

If the parents are non-inbred and unrelated, then a pair of siblings is expected to share 0, 1, or 2 alleles IBD with probabilities ¼, ½, and ¼ respectively. However, if both of the siblings are affected, they are expected to share more alleles IBD at loci linked to a disease gene. Affected-sib-pair IBD-based linkage methods test for departures of IBD sharing from the expectations under the null hypothesis of no linkage, in a sample of ASPs.

Using multipoint IBD sharing estimates, GENEHUNTER computes a commonly used model-free linkage statistic, known as the NPL (non-parametric linkage) statistic. Very generally, the NPL statistic measures the amount of excess IBD sharing between affected individuals in a sample of families. An example of a NPL plot is shown in Figure 1.4. A large NPL value at a locus provides evidence of linkage between that locus and a disease gene. Although multipoint methods can be used in the estimation of marker IBD sharing, which is then used to calculate the NPL score at a given locus, testing for linkage is still done at each locus separately. Independently testing for linkage at a series of locations, which is

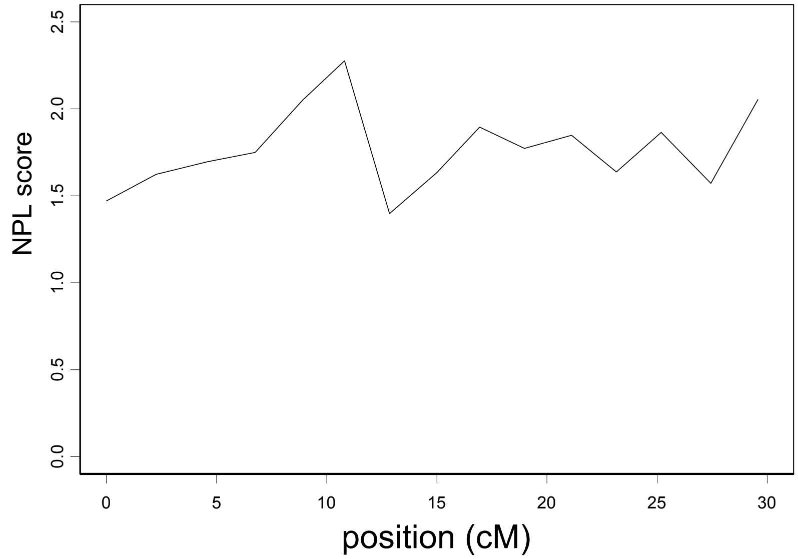


Figure 1.4: Example of a NPL plot

common to many linkage methods, gives rise to multiple testing problems. If the NPL peak is sufficiently high, the location of the peak is usually taken as a point estimate of a disease gene location. However, it is clear that the NPL method is primarily designed as a testing method (testing for linkage), rather than an estimation method (for estimating the location of a disease gene). Since linkage genome scans are usually only the first step in the search for disease susceptibility genes, identification of regions that may harbor such genes is their main goal. Regions that are likely to contain disease genes can then be further analyzed in fine-mapping studies. Although allele-sharing linkage methods are widely used for the study of complex traits, most of these methods are not designed to estimate the location(s) of disease gene(s) with a specified level of certainty, and they do not provide confidence intervals for disease gene locations. Although several authors have suggested approaches for constructing confidence regions (see Kruglyak and Lander 1995; Lin 2002; and Hössjer 2003), they have not been widely applied.

1.2.2 Estimation of gene location and effect using a GEE approach

Recently, Liang et al. (2001*a*) proposed analyzing marker IBD sharing data from a sample of affected sib pairs using generalized estimating equations (GEE) to estimate the location of a trait gene within a chromosomal region framed by multiple markers, as well as the effect size of the gene, measured in terms of expected allele-sharing among affected sibs. It is well known that for affected relative pairs, deviations from the null expectations for marker allele sharing depend on mode-of-inheritance parameters and the number of loci involved, as well as on the recombination fraction between the marker and the disease locus

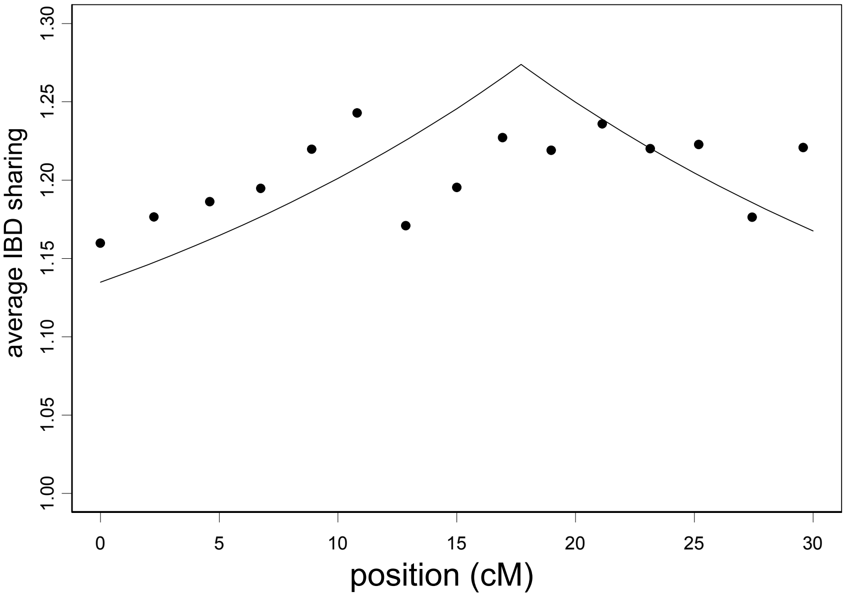


Figure 1.5: Example of an expected IBD sharing curve fit using the GEE method. (The points are the average marker IBD sharing in a sample of affected sib pairs.)

(Risch 1990). Liang et al. (2001*a*) derived an expression for the expected allele sharing among affected sib-pairs at any position in a chromosomal region, in terms of the expected allele sharing at a disease locus in that region and the recombination fraction between that position and the disease locus. They used this expression as a basis for estimating the location of the disease gene and the expected allele sharing at that gene. A key feature of this method is that it uses data from all the linked markers simultaneously to fit the expected IBD sharing curve to IBD sharing data from a sample of ASPs, thereby providing estimates of the location of the disease gene and expected IBD sharing in ASPs at that gene. An example of an expected IBD sharing curve, estimated by this method for a particular data set, is shown in Figure 1.5. The method will be described in detail in section 2.4.

This approach overcomes some of the shortcomings of other allele-sharing based methods. Using data at all markers simultaneously, can presumably improve accuracy and precision of gene location estimates. A confidence interval for the map position of the susceptibility gene can be constructed using this method, which is useful for defining regions for further fine-mapping studies. Furthermore, multiple testing for linkage to many linked markers is avoided. These advantages come at a cost: the method relies on the explicit assumption that there is exactly one disease gene in the region, and is not robust to violations of this assumption.

Liang et al. (2001*a*) pointed out that their GEE approach is not meant as a substitute to currently used linkage methods such as the NPL. Rather, they propose it as an adjunct to genome scanning approaches. Given (suggestive) evidence of linkage from a standard linkage genome scan, they propose applying their method to refine the estimate of the disease gene location and assess evidence for linkage at that position.

1.3 Mapping Genes Contributing to Complex Diseases

Recent research in human genetics has focused largely on investigating traits controlled by multiple genetic and environmental factors, as well as their interactions. These so-called complex traits, such as diabetes, hypertension and many psychiatric disorders, are a major public health concern, and an improved understanding of their etiology can potentially result in substantial health care benefits. Standard linkage methods have not been very successful in mapping genes that contribute to these traits. Primarily, this is because each individual gene may have relatively small impact on disease susceptibility, leading to reduced power of linkage methods.

Methods that take into account the role of multiple genes and environmental factors may be more powerful and improve the precision of disease-gene localization in the analysis of genetic data for complex traits. This idea has led to increased interest in methods that incorporate the fact that there may be more than one gene contributing to the disease. A review of research concerning such methods will be presented in section 2.1. The literature review revealed that most "two-locus" methods were designed primarily to test for linkage to multiple susceptibility genes, rather than to localize multiple disease genes. Two-locus methods that can be used to map a second interacting gene, generally require the first disease gene to have already been mapped (Cordell et al. 1995; Farrall 1997; Cox et al. 1999; Cordell et al. 2000; Strauch et al. 2000; Liang et al. 2001*b*; Holmans 2002; Chiu and Liang 2004). In addition, the majority of methods that attempt to take into account the existence of two disease genes assume that the two putative genes are unlinked to one another.

1.4 Motivation for Thesis

As was mentioned in the previous section, one drawback of most of the two-locus linkage methods described in the literature is that they are designed to deal only with two unlinked disease genes. The few existing methods designed for the study of the effects of two linked disease genes (Delépine et al. 1997; Farrall 1997) can be classified as "conditional methods", in that they require prior localization of a primary gene, and subsequently assess evidence for a second putative gene, conditional on the effect of the first gene with a "known" location. However, the location of the first locus that is being conditioned on may be inaccurate if it was mapped without taking into account the existence of linked trait genes. Conditioning on this inaccurate position may further lead to incorrect results for the second gene. Methods for simultaneously mapping two linked disease genes are likely to improve the accuracy of localization of both genes.

The existence of linked genes contributing to a single trait appears to be plausible in complex diseases (e.g. Luo et al. 1996; Delépine et al. 1997; Ghosh et al. 1999; Hampe et al. 2002). Evidence suggesting this comes both from animal model and human disease studies. Non-parametric linkage genome scans can yield wide peaks, or multiple peaks within the same region. Such observations could be the result of two or more susceptibility genes for the same disease in one chromosomal region. Conventional linkage analysis can be misleading in such situations, and the peak linkage score may not occur at either trait gene location (Hauser et al. 2003). We extended the methods of Liang et al. (2001*a*) to localize two linked disease susceptibility genes. Our method is designed to estimate the locations of the two genes simultaneously using all marker data in a region, leading to more accurate and more precise localization. Furthermore, use of the GEE approach allows construction of confidence intervals for disease gene locations, which may be useful for defining regions for fine-mapping studies.

1.5 Thesis Summary

In chapter 2 a more detailed review of relevant literature is provided and the method proposed by Liang et al. (2001*a*) for estimation of a single disease gene location by a GEE approach is described. In chapter 3 several extensions and modifications of this method are discussed. In section 3.1 alternative formulations of the variance function in the implementation of a GEE approach for estimation of a single disease gene location are considered. In section 3.2 a different working correlation matrix is investigated. In section 3.3 an alternative parameterization of the problem is presented, which allows the estimation of additional parameters from the genetic disease model and may improve estimation of the disease gene location. In chapter 4 a method of estimating the locations and expected allele sharing at two linked disease genes is derived and its implementation is described. Chapter 5 contains a summary of a simulation study carried out to assess properties of the GEE estimation method proposed in chapter 4. In chapter 6 we introduce test statistics to help determine the number of disease genes in a single chromosomal region. Properties of these test statistics are evaluated by simulation in chapter 7. In chapter 8, an application of the methods described in this thesis to type 1 diabetes data is presented. The main findings from this thesis are summarized in chapter 9.

Chapter 2

Literature Review: Statistics and Statistical Genetics Background

2.1 Statistical Methods for the Study of Multi-locus Disease Systems

For complex diseases, methods that take into account the role of multiple genes and environmental factors may be more powerful and improve the precision of disease-gene localization. Several analysis methods that take into account existence of multiple (usually two) disease genes have been developed. Both "parametric" (e.g. Goldin and Weeks 1993; MacLean et al. 1993; and Schork et al. 1993) and "nonparametric" (e.g. Knapp et al. 1994; Cordell et al. 1995; Cox et al. 1999; Cordell et al. 2000) approaches to modeling the effects of multiple loci have been considered. The main concerns with parametric methods are that they require specification of many unknown parameters, and they tend to be computationally intensive.

When developing and assessing multi-locus linkage methods, it is important to identify the main purpose of the analysis. The primary purpose of a study may be the detection or localization of new disease loci, perhaps taking into account "known" disease loci. Alternatively, the primary purpose may be to resolve the nature of interactions among a number of loci. Schork et al. (1993) comment that they do not expect "2-trait-locus 2-marker-locus analyses" to be used in initial genome scans. They believe these methods would be more useful for a set of candidate loci, or in situations where preliminary evidence of linkage exists. On the other hand, Cordell et al. (1995) propose using their two-locus statistics as the basis of a semi-simultaneous search procedure. Cordell et al. (2000) state that the primary aim of their methods is the detection of disease loci in the presence of epistasis, rather than modeling the interaction.

The benefits of a two-locus or multilocus method can be measured in various ways, including the magnitude of the test statistic (e.g. expected maximum lod score), power, or accuracy of gene localization. The potential increase in power of linkage analysis resulting from the simultaneous consideration of two susceptibility genes has been assessed for various methods (e.g. Goldin and Weeks 1993; Knapp et al. 1994; Sham et al. 1994; Cordell et al. 1995; Zinn-Justin and Abel 1998; Cordell et al. 2000; Farrall 1997; Cox et al. 1999; Lin 2000; Holmans 2002). It has been demonstrated that allowing for different modes of interaction between potential disease loci can lead to increased power to detect any one of the loci, however, this increase in power may be modest (Cordell et al. 2001).

Several papers have described methods that can be used to study interactions between two susceptibility genes (e.g. Cordell et al. 1995; Tiwari and Elston 1997; Cox et al. 1999; Cordell et al. 2001). MacLean et al. (1993) and Cox et al. (1999) proposed using single-locus statistics for inferences about interactions, and found that correlation between single-locus statistics can be useful for detecting multi-locus systems, however, not all types of interaction can be detected by correlation. In order to study the interactions between genes, one must first identify the genes involved by finding their positions in the genome. Thus precise localization of disease susceptibility genes is of primary importance.

Schork et al. (1993) compared accuracy of estimates of recombination fractions, and found that they were more accurate for their analyses that took into account the effects of two trait loci. Many of the early two-disease-gene methods designed for discovery of linkage or for studying interactions between loci did not use a map of markers, but rather used only a single marker or pair of unlinked markers (Schork et al. 1993; Knapp et al. 1994; Tiwari and Elston 1997; Zinn-Justin and Abel 1997). Thus, most of these methods are not well suited to localizing disease genes. Even those two-locus methods that can be used to map a second interacting gene, generally require the first disease gene to have already been mapped (Cordell et al. 1995; Farrall 1997; Cox et al. 1999; Cordell et al. 2000; Strauch et al. 2000; Liang et al. 2001*b*; Holmans 2002; Chiu and Liang 2004). Methods for simultaneously mapping two disease genes are likely to improve the precision of gene localization.

A major drawback of most of the two-locus linkage methods described in the literature is that they are designed to deal only with unlinked disease genes. The problem of two linked loci has not been examined extensively in the literature, excepting the work of Delépine (1997), Farrall (1997), Cordell et al. (1998) and Biswas et al. (2003). Cordell et al. (1998) described association methods for mapping multiple linked QTLs in inbred animal populations. The focus of Farrall's (1997) work was to resolve genetic interactions between linked loci and improve power to detect linkage to a second gene, by taking into account evidence for linkage at a linked gene with an established location. Their results showed that the power to detect linkage was maintained, provided the two susceptibility genes were not tightly linked. They suggested using their methods to assess independent support for putative susceptibility genes that map close to established genes, or to analyze “ill-defined” or “broad” peaks that may be due to multiple disease genes. This approach was applied to IDDM data by Cordell et al. (2000), showing some evidence for a second type 1 diabetes gene on chromosome 6 linked to IDDM1. Motivated by an analysis of type 1 diabetes data, Delépine et al. (1997) also developed a method to test for the presence of a second susceptibility gene linked to a known first susceptibility gene. Their method applies weighted logistic analysis to fully informative IBD sharing data at two markers to test for evidence of linkage to a second marker conditional on linkage to a first marker, rather than mapping the two disease genes simultaneously. No simulation study assessing the performance of this method was reported. For both Delépine et al. (1997) and Farrall’s (1997) methods, the localization of the first locus that is being conditioned on could be inaccurate if it was mapped without taking into account the existence of linked trait genes. Conditioning on an inaccurate position of the first gene may lead to incorrect results regarding the second gene. Biswas et al. (2003) applied a Bayesian approach that accounts for heterogeneity to the simultaneous detection of two linked disease genes. This method, however, is a parametric approach requiring the specification of the penetrance vectors, designed to detect genes under locus heterogeneity. The method provides point estimates of the two disease gene locations, but no confidence intervals for these locations were reported.

2.2 Generalized Estimating Equations Methods

In ASP-based model-free genome scans, we consider IBD sharing of sibling pairs at a series of markers, and test for deviations from the expectation under the null hypothesis of no linkage to a disease gene at each marker. It is well known, that for a pair of relatives, regardless of the mode of inheritance, IBD sharing at two linked markers is correlated due to transmission of chromosomal fragments. Therefore, the outcome of interest in IBD-based genome scans, i.e. the IBD sharing at a series of linked markers for a sample of affected sibling pairs, is comprised of independent clusters, i.e. sibpairs, with correlated outcomes, i.e. marker IBD allele sharing, within clusters. A similar data structure is observed in, for example, longitudinal studies, where multiple observations are made on each individual over a period of time. While observations made on different individuals are independent, observations made on the same individual at different times are correlated. In an allele-sharing genome scan, marker map position is analogous to a time covariate in longitudinal data. To obtain valid inference, correlation among values for a given subject should be taken into account during analysis of such data. Liang and Zeger (1986) proposed the use of generalized estimating equations (GEEs) for longitudinal data analysis. In this section some general concepts and features of the GEE approach will be described.

The GEE approach can be viewed as an extension of quasi-likelihood to the analysis of dependent data (Zeger and Liang 1986). With the quasi-likelihood approach (Wedderburn 1974; McCullagh 1983), parameters $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^{\mathrm{T}}$ are estimated by solving the quasi-score estimating equations:

$$\mathbf{U}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{D}^{T}\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})/\sigma^{2} = \mathbf{0} \qquad (2.1)$$

where $\mathbf{Y} = (Y_1,\ldots,Y_n)^{\mathrm{T}}$ is a vector of observations, $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}) = (\mu_1,\ldots, \mu_n)^{\mathrm{T}}$ and $\mathrm{cov}(\mathbf{Y}) = \sigma^2\mathbf{V}(\boldsymbol{\mu})$.The parameters of interest, $\boldsymbol{\beta}$, relate to the dependence of $\boldsymbol{\mu}$ on the covariates, so $\boldsymbol{\mu}$ is a function of the parameters of interest, i.e. $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\beta})$. The covariance of the data vector, $\mathrm{cov}(\mathbf{Y}) = \sigma^2\mathbf{V}(\boldsymbol{\mu})$, is assumed to be a function of the mean, $\boldsymbol{\mu}$, and potentially a scale parameter, σ. The components of the $n\times p$ matrix $\mathbf{D}$ are $\mathbf{D}_{ir} = \partial\mu_i/\partial\beta_r$. It is evident from the above estimating equations that quasi-likelihood estimation does not require specification of the likelihood. Instead, only the mean, $\boldsymbol{\mu}$ and the variance $\sigma^2\mathbf{V}(\boldsymbol{\mu})$ need to be specified.

If a data vector $\mathbf{Y}^{\mathrm{T}} = (\mathbf{Y}_1^{\mathrm{T}},\ldots, \mathbf{Y}_K^{\mathrm{T}})$ consists of sets of correlated observations from K independent individuals, the covariance matrix of $\mathbf{Y}$, will be block-diagonal. Block-diagonal covariance matrices arise frequently, for example in longitudinal data. The correlation between data within blocks is often unknown and difficult to specify. Liang and Zeger (1986) and Zeger and Liang (1986) suggested the use of generalized estimating equations in such situations. Before describing the GEE approach, some notation more suitable for discussion of this type of block-data will be introduced. With correlated data, for each cluster or subject i, there are several measurements of a response $Y_{ij}, j=1,\ldots,n_i$. [If $n_i=n$ for all i, then $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{in})$ for $i=1,\ldots,K$.] Assume $\mathrm{E}(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ and $\mathrm{cov}(\mathbf{Y}_i) = \mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i\mathbf{A}_i^{\frac{1}{2}}$ where $\mathbf{A}_i$ is a $n_i\times n_i$ diagonal matrix with $\mathrm{var}(Y_{ij})\, j=1,\ldots, n_i$ on the diagonal, and $\mathbf{R}_i$ is a matrix of correlations of elements of $\mathbf{Y}_i$. Also let $\mathbf{D}_i$ be the $n_i\times p$ matrix of derivatives of the mean function for the ith subject with respect to the parameters, i.e. $\mathbf{D}_i(m,r) = \partial\mu_{im}/\partial\beta_r$.

In this case, parameter estimates, $\hat{\boldsymbol{\beta}}$, are also obtained by solving the estimating equations (2.1) which for this type of block-structured data can be written as a sum over the K independent subjects. Liang and Zeger (1986) explain that the covariance of the data for the ith subject, $\mathrm{cov}(\mathbf{Y}_i) = \mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i\mathbf{A}_i^{\frac{1}{2}}$, can be replaced by $\mathbf{V}_i = \mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_{0i}(\alpha)\mathbf{A}_i^{\frac{1}{2}}$ where $\mathbf{R}_{0i}(\alpha)$ is referred to as the "working" correlation, and may depend on a set of parameters α. This formulation leads to the generalized estimating equations:

$$\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}\,. \tag{2.2}$$

An attractive feature of the GEE approach is that obtaining a consistent estimator for β does not require $\mathbf{R}_{0i}(\alpha)$ to be specified correctly. The consequence of an incorrectly specified working correlation is reduced efficiency of the estimator. Likewise, it is not necessary in this approach to specify var(Y_{ij}) = $(\mathbf{A}_i)_{jj}$ correctly. Under mild regularity conditions, Liang and Zeger (1986) show that as $K \to \infty$, $\hat{\boldsymbol{\beta}}$ obtained by solving the estimating equations (2.2) is a consistent estimator of β and is asymptotically multivariate normal with covariance matrix $\mathbf{V}_R$ given by:

$$\mathbf{V}_R = \left(\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \left[\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \operatorname{cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i \right] \left(\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \tag{2.3}$$

(Huber 1967, White 1982). In the above, cov($\mathbf{Y}_i$) is the actual covariance of $\mathbf{Y}_i$, rather than the assumed covariance under the "working correlation" $\mathbf{R}_{0i}(\alpha)$. The estimator $\mathbf{V}_R$ is known as the robust, or sandwich, variance estimator. To estimate $\mathbf{V}_R$, cov($\mathbf{Y}_i$) can be approximated by $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}$ and β can be approximated by $\hat{\boldsymbol{\beta}}$. Note that if the working correlation for the data has been correctly specified so that cov($\mathbf{Y}_i$) = $\mathbf{V}_i$, then the above estimator reduces to:

$$\mathbf{V}_N = \left(\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} . \tag{2.4}$$

This estimator is known as the naïve variance estimator.

Assume that $\mathbf{R}_{0i}(\alpha) = \mathbf{R}_0(\alpha)$ for all i. Some simple correlation structures that are frequently assumed are the independence correlation structure where $\mathbf{R}_0 = \mathbf{I}$ (the identity matrix), the exchangeable correlation matrix where $[\mathbf{R}_0]_{jk} = \alpha$ for all $j \neq k$, or the first-order autoregressive correlation where $[\mathbf{R}_0]_{jk} = \alpha^{|k-j|}$ for all $j \neq k$. An advantage of the independence assumption is that the estimator is easier to compute and it may be reasonably efficient for a few simple designs. However, using an overly simple working correlation may yield less efficient estimates in cases where the correlation is large (Liang and Zeger 1986).

Parameter estimates and estimates of their variances provided by the GEE approach are typically used to construct Wald Chi-square test statistics and asymptotic confidence intervals. However, in small samples, the sandwich estimator, $\mathbf{V}_R$, has been shown to result in downwardly biased variance estimates (e.g. Pan 2001). Drum and McCullagh (1993) discussed the fact that the sandwich estimator may not work well in small samples. A Wald test based on approximating cov($\hat{\boldsymbol{\beta}}$) by $\hat{\mathbf{V}}_R$ ignores both the bias and variation of the variance estimator. Methods have been proposed to correct for the bias (e.g. Mancl and DeRouen 2001; Pan 2001) and the variability (e.g. Mancl and DeRouen 2001; Pan and Wall 2002) of the robust variance estimator.

In summary, the key features of the GEE approach are that it does not require a fully specified likelihood, and the asymptotic validity of the GEE estimates depends only on the correct specification of the mean function of the data, $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{Y}_i)$. Generalized estimating equations allow one to estimate parameters with greater efficiency by accounting for correlation between observations, rather than simply assuming an independence correlation structure, while still obtaining valid estimates of the variances of parameter estimates via the sandwich estimator.

2.3 Hypothesis Testing in the GEE Framework

2.3.1 Background: Wald, score and likelihood ratio tests in the maximum likelihood estimation framework

Assume θ is an unknown parameter vector of length p. Let L(θ|y) = f(y;θ) be the likelihood function for the parameter θ, and $\hat{\theta}$ be the maximum likelihood estimate (MLE) of θ. That is, $\hat{\theta}$ is the value of θ that maximizes the likelihood function. The function S(θ) = ∂logL/$\partial\theta$ is referred to as the score function, and the MLE $\theta = \hat{\theta}$ solves the equations S(θ) = ∂logL/$\partial\theta$ = 0. The asymptotic covariance matrix of $\hat{\theta}$ is I^{-1} where $I = -E\left(\frac{\partial^2 \log \mathrm{L}}{\partial \theta^2}\right)$ is the Fisher information matrix for θ.

The Wald, score, and likelihood-ratio test statistics for testing H$_0$: $\theta = \theta_0$ vs H$_A$: $\theta \neq \theta_0$ can be calculated as:

$$\chi^2_{wald} = \left(\hat{\theta} - \theta_0\right)' I(\hat{\theta})\left(\hat{\theta} - \theta_0\right),$$

$$\chi^2_{score} = S(\theta_0)' I^{-1}(\theta_0) S(\theta_0), \text{ and}$$

$$\chi^2_{LR} = -2\left\{\log \mathrm{L}\left(\theta_0 \mid y\right) - \log \mathrm{L}\left(\hat{\theta} \mid y\right)\right\}.$$

Asymptotically, under the null hypothesis, all three of the above test statistics follow a chi-squared distribution with p degrees of freedom.

More generally one may want to test hypotheses that restrict a certain subset of parameters. In other words, we may want to test a hypothesis which specifies a sub-model of the fully parameterized model. Let $\hat{\theta}_0$ and $\hat{\theta}$ be the MLEs of θ under the null hypothesis or sub-model and the unrestricted model, respectively. Then asymptotically, under the null hypothesis, the likelihood ratio statistic and score statistic given by:

$$\chi^2_{LR} = -2\left\{\log \mathrm{L}\left(\hat{\theta}_0 \mid y\right) - \log \mathrm{L}\left(\hat{\theta} \mid y\right)\right\} \text{ and}$$

$$\chi^2_{score} = S\left(\hat{\theta}_0\right)' I^{-1}\left(\hat{\theta}_0\right) S\left(\hat{\theta}_0\right)$$

are both approximately chi-squared with degrees of freedom equal to the difference between the numbers of parameters specified under the sub-model and the unrestricted model. Now assume that the hypotheses about a subset of parameters can be formulated as H$_0$: $\mathbf{C}\theta = (\mathbf{C}\theta)_0$ vs H$_A$: $\mathbf{C}\theta \neq (\mathbf{C}\theta)_0$ where $\mathbf{C}$ is a matrix of dimension r by p. Then the Wald statistic for testing these hypotheses is given by:

$$\chi^2_{wald} = \left(\mathbf{C}\hat{\theta} - \mathbf{C}\theta_0\right)' \mathrm{Cov}^{-1}\left(\mathbf{C}\hat{\theta}\right)\left(\mathbf{C}\hat{\theta} - \mathbf{C}\theta_0\right) = \left(\mathbf{C}\hat{\theta} - \mathbf{C}\theta_0\right)' \left\{\mathbf{C}I^{-1}\mathbf{C}'\right\}^{-1}\left(\mathbf{C}\hat{\theta} - \mathbf{C}\theta_0\right),$$

and has an asymptotic χ^2_r distribution under H$_0$.

Further, suppose $\theta' = (\theta_1', \theta_2')$ where θ_1 is a vector of r parameters of interest and θ_2 is vector of p-r nuisance parameters, and we are interested in testing the hypothesis H$_0$: $\theta_1 = \theta_{10}$. If the null hypothesis can be stated in this way, as a simple hypothesis for a subset of parameters, the score test statistic can also be calculated as follows. Partition the score vector and matrix I in a way corresponding to the partition of θ, and let $\hat{\theta}_{20}$ be the MLE of θ_2 given $\theta_1 = \theta_{10}$, and $\hat{\theta}_0' = \left(\theta_{10}', \hat{\theta}_{20}'\right)$. Then the score statistic is:

$$\chi^2_{1.2} = S_1\left(\hat{\theta}_0\right)' \left\{I_{11.2}\left(\hat{\theta}_0\right)\right\}^{-1} S_1\left(\hat{\theta}_0\right)$$

where $I_{11.2} = I_{11} - I_{12}I_{22}^{-1}I_{12}'$ (I_{ij} represents a submatrix of I). Under H$_0$, this statistic is asymptotically a χ^2_r random variable.

2.3.2 Tests in the GEE/quasi-likelihood framework

The generalized estimating equations approach for parameter estimation was described in section 2.2. This approach only requires specification of the mean and covariance functions, without specifying the entire distribution of the observations. Since in a GEE framework a likelihood function is not defined and maximum-likelihood estimates are not calculated, the Wald, score, and likelihood ratio tests described in section 2.3.1 cannot be applied directly. However, there are analogous tests which can be applied in the GEE or quasi-likelihood framework.

With the GEE approach, estimates of a parameter $\boldsymbol{\beta}$ are obtained by solving the generalized estimating equations. Consistent estimates of the parameters can be obtained if the mean function is correctly specified, even if the covariance matrix for the data is misspecified (Liang and Zeger 1986). Furthermore, the solution of the GEEs is asymptotically normally distributed with mean $\boldsymbol{\beta}$. If the covariance matrix of the data is correctly specified, then the covariance matrix of the parameter estimate is given by the model-based (or naïve) variance estimator (equation 2.4); whereas, if the covariance matrix of the data may be misspecified, covariance of $\hat{\boldsymbol{\beta}}$ can be estimated by the robust variance estimator $\mathbf{V}_R$ (equation 2.3)

GEE Wald Test:

Since the GEE estimate of $\boldsymbol{\beta}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}_R$ (equation 2.3), Wald tests can be formulated using the GEE estimate $\hat{\boldsymbol{\beta}}$ and the robust variance estimate $\mathbf{V}_R$. Rotnizky and Jewell (1990) described extensions of the usual chi-squared tests for testing hypotheses about a subset of regression parameters for cluster correlated data. Let $\hat{\boldsymbol{\beta}}$ be the GEE estimate of $\boldsymbol{\beta}$, obtained by solving the equations

$$\mathrm{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i{}' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \text{ where } \mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}},\ \mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_{0i}(\alpha)\mathbf{A}_i^{1/2} \text{ with}$$

$\mathbf{A}_i$=diag{var(y_{ij})} and $\mathbf{R}_{0i}$ being the working correlation. Again consider partitioning the regression parameter vector such that $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ where $\boldsymbol{\beta}_1$ is r by 1 and $\boldsymbol{\beta}_2$ is p-r by 1, $r \leq p$, and assume we are interested in testing the hypothesis H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$. Because of the asymptotic normality of $\hat{\boldsymbol{\beta}}$, the generalized Wald test statistic:

$$T_w = \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\right)' \hat{\mathbf{V}}_{\beta_1}^{-1} \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\right) \tag{2.5}$$

has a χ_r^2 distribution under H_0. $\hat{\mathbf{V}}_{\beta_1}$ is the r by r sub-matrix of $\hat{\mathbf{V}}_R$, comprised of elements corresponding to the elements of $\boldsymbol{\beta}_1$.

GEE Score Test:

Rotnizky and Jewell (1990) define the generalized score test statistic as:

$$T_S = \mathbf{U}_{\beta_1}\left(\boldsymbol{\beta}_{10}, \tilde{\boldsymbol{\beta}}_{20}\right)' \tilde{\Sigma}_{\beta_1}^{-1} \mathbf{U}_{\beta_1}\left(\boldsymbol{\beta}_{10}, \tilde{\boldsymbol{\beta}}_{20}\right) \tag{2.6}$$

where $\tilde{\boldsymbol{\beta}}_{20}$ is the GEE estimator of $\boldsymbol{\beta}_2$ in the restricted parameter space under H_0,

$$\mathbf{U}_{\beta_1} = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_1} \right)' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \text{, and}$$

$$\Sigma_{\beta_1} = \mathbf{W}_{\beta_1}^{-1} \mathbf{V}_{\beta_1} \mathbf{W}_{\beta_1}^{-1} \tag{2.7}$$

with $\mathbf{W}_{\beta_1}$ being a sub-matrix of $\mathbf{W} = \left\{ \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right\}^{-1}$.

Breslow (1990) provides a computationally simpler formula for estimating Σ_{β_1}:

$$B_{11} - A_{12} A_{22}^{-1} B_{21} - B_{12} A_{22}^{-1} A_{21} + A_{12} A_{22}^{-1} B_{22} A_{22}^{-1} A_{21} \tag{2.8}$$

where $B = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)' \mathbf{V}_i^{-1} \mathbf{D}_i$, $A = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i$, and A_{ij} and B_{ij} are sub-matrices of A and B.

Approximate Quasi-Likelihood Ratio Tests:

Given the score function, $S(\theta)$, for a likelihood $\mathrm{L}(\theta)$, the likelihood ratio $\mathrm{L}(\eta)/\mathrm{L}(\theta)$ is given by:

$$\lambda(\theta, \eta) = \frac{\mathrm{L}(\eta)}{\mathrm{L}(\theta)} = \exp\left(\int_{\theta}^{\eta} S(t)\, dt \right).$$

By analogy, given the quasi-score function, $\mathrm{QS}(\theta)$, the quasi-likelihood ratio is defined as:

$$\mathrm{QLR}(\theta, \eta) = \exp\left(\int_{\theta}^{\eta} \mathrm{QS}(t)\, dt \right).$$

McCullagh and Nelder (1989) point out that when θ is a vector-valued parameter, definition of the quasi-likelihood is problematic. The most direct analogy to the above would be a line integral between θ and η, however, generally, this line integral would be path-dependent, and therefore, the quasi-log-likelihood is not uniquely defined. Although the quasi-score can still be used to construct confidence sets even in the absence of the quasi-log-likelihood, it has undesirable properties when the quasi-score has multiple roots (McCullagh 1991; McLeish and Small 1992).

McLeish and Small (1992) introduce a class of inference functions for independent observations that contain analogues of likelihood ratios under semiparametric assumptions. They consider the projection of the likelihood ratio into this class of inference functions and show that this projected likelihood is first order equivalent to the quasi-likelihood. Based on their projected likelihood, McLeish and Small (1992) propose approximating the likelihood ratio by:

$$\hat{\lambda}(\theta,\eta)=\prod_{i=1}^{n}\left[1+\left(\mu_i(\eta)-\mu_i(\theta)\right)\left(Y_i-\mu_i(\theta)\right)/\sigma_i^2(\theta)\right]$$

where $(Y_1,\ldots,Y_n)$ is the data vector, $\mu_i = \mathrm{E}(Y_i)$ and $\sigma_i^2 = \mathrm{Var}(Y_i)$, and describe how it can be used to test statistical hypotheses for independent random variables. For multivariate observations they suggest the following extension of their quasi-likelihood ratio which uses the covariance matrix of $\mathbf{Y}_i$, Σ_i:

$$\hat{\lambda}(\theta,\eta)=\prod_{i=1}^{n}\left[1+\left\{\boldsymbol{\mu}_i(\eta)-\boldsymbol{\mu}_i(\theta)\right\}^T\boldsymbol{\Sigma}_i^{-1}(\theta)\left\{\mathbf{Y}_i-\boldsymbol{\mu}_i(\theta)\right\}\right] \tag{2.9}$$

where $\mathbf{Y}_i$ is the ith vector-valued (multivariate) observation, $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{Y}_i)$, and $\boldsymbol{\Sigma}_i = \mathrm{Cov}(\mathbf{Y}_i)$.

Li (1993) introduced a linear deviance function that can be used in conjunction with the quasi-likelihood method and is defined for both independent and dependent observations. He proposes approximating the quasi-log-likelihood ratio by

$$\hat{R}(\eta,\theta,\mathbf{Y})=\frac{1}{2}\left(\boldsymbol{\mu}_\eta-\boldsymbol{\mu}_\theta\right)^T\boldsymbol{\Sigma}_i^{-1}(\theta)\left(\mathbf{Y}-\boldsymbol{\mu}_\theta\right)+\frac{1}{2}\left(\boldsymbol{\mu}_\eta-\boldsymbol{\mu}_\theta\right)^T\boldsymbol{\Sigma}_i^{-1}(\eta)\left(\mathbf{Y}-\boldsymbol{\mu}_\eta\right) \tag{2.10}$$

where $\mathbf{Y}^T = (\mathbf{Y}_1^T,\ldots, \mathbf{Y}_n^T)$, $\boldsymbol{\mu} = E(\mathbf{Y})$, and $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{Y})$.

In certain situations, when the quasi-score has multiple roots, the confidence set based on this statistic is better than that based on the score test. Li (1993) notes that $\hat{R}(\eta,\theta,X)$ is a useful alternative to the method of McLeish and Small (1992) partly because it applies to dependent observations. Li also comments that if the quasi-score has multiple roots, $\hat{R}(\eta,\theta,X)$ can be used to distinguish the correct solution from incorrect ones.

2.4 Estimation of a Single Disease Gene Location and Expected IBD Sharing Using a GEE Approach

Recently an IBD-based procedure to estimate the location of a susceptibility gene within a chromosomal region framed by multiple markers was proposed by Liang et al. (2001*a*). The authors noted that the main purpose of the procedure is to estimate the location of a susceptibility gene when there is preliminary evidence that the chromosomal region includes a disease gene. Liang et al. (2001*a*) derived a simple representation relating the expected IBD sharing at a single marker to its distance from the disease locus. The authors first proposed exploratory analysis for locating the disease gene. They then developed a formal procedure to infer point and interval estimates of the location of the disease gene by applying the GEE approach to the problem.

The proposed procedure requires no assumptions about the mode of inheritance, except that the region contains no more than one susceptibility gene. Under the assumptions of random mating, linkage equilibrium, and generalized single ascertainment, Liang et al. derived a simple representation relating the mean IBD sharing in affected sib pairs (ASPs) at any position, *t*, to their mean sharing at the disease locus, and the distance between *t* and the disease locus. Assuming a region with one susceptibility gene at location τ, they showed that:

$$\mu_1(t) = E(S(t) \mid \Phi) = 1 + (2\Psi_{t,\tau} - 1)(E(S(\tau) \mid \Phi) - 1)$$

where $S(t)$ = number of alleles (0, 1, or 2) shared IBD by an ASP at locus *t* and Φ denotes the event that both sibs are affected. The subscript 1 on the mean function μ signifies that this is the expected sharing assuming one disease gene in the region, i.e. under the one-locus model. The above expression for $\mu_1(t)$ is robust, in the sense that it applies regardless of the true mode of inheritance. The map distance between *t* and τ is measured by:

$$\Psi_{t,\tau} = \Psi(\theta_{t,\tau}) = \theta_{t,\tau}^2 + (1 - \theta_{t,\tau})^2,$$

where $\theta_{t,\tau}$ is the recombination fraction between *t* and τ. Using Haldane's map function to relate recombination fraction to genetic distance leads to:

$$\mu_1(t) = E(S(t) \mid \Phi) = 1 + e^{-.04|t-\tau|}C,$$

where $C = E(S(\tau) \mid \Phi) - 1$ (i.e. one less than the expected number of alleles shared IBD at τ by an ASP). It is easily seen that $\mu_1(t)$ is strictly decreasing in $|t\text{-}\tau|$ and attains its maximum at τ.

Liang et al. (2001*a*) developed an inferential procedure to estimate the location of the disease gene, τ, and $C = E(S(\tau) \mid \Phi) - 1$ using marker IBD sharing data from ASPs. Assume an ASP design with *M* markers at positions $t_1,...,t_M$. For each sib-pair $Y_i = (Y_i(t_1),\ldots,Y_i(t_M))$ is observed, where $Y_i(t_j)$ represents all the marker information at locus t_j (j=1,…,M) for the *i*th pedigree, $i = 1,\ldots,n$. In a situation with fully informative markers, $S_i(t_j)$ could be counted directly and $\mu_1(t_j)$ could be estimated by:

$$\bar{S}(t_j) = \sum_{i=1}^{n} S_i(t_j)/n$$

If the *j*th marker is not fully informative for family *i*, $S_i(t_j)$ is unknown, but can be estimated by considering all possible IBD configurations consistent with the observed marker information, i.e. $S_i(t_j)$ can be imputed given Y_i by:

$$S_i^*(t_j) = \sum_{l=0}^{2} l \cdot \Pr(S_i(t_j) = l \mid Y_i).$$

In such situations, $\mu_1(t_j)$ can be estimated by:

$$\bar{S}^*(t_j) = \sum_{i=1}^{n} S_i^*(t_j)/n$$

Therefore, Liang et al. (2001*a*) propose using the statistics $S_i^*(t_j)$ as the basis for estimating $\delta_1 = (\tau,C)$. They estimate the parameters of the one-locus model $\delta_1 = (\tau,C)$ by solving the following estimating equations:

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_1(\delta_1)}{\partial \delta_1} \right)' \mathrm{Cov}^{-1}(S_i^* \mid \Phi)(S_i^* - \mu_1(\delta_1)) = 0$$

where $S_i^* = (S_i^*(t_1),\ldots, S_i^*(t_M))'$, and $\mu_1(\delta_1) = (\mu_1(t_1;\delta_1),\ldots, \mu_1(t_M;\delta_1))'$. Estimates for τ and C and their estimated precision are consistent as long as $\mathrm{E}[S_i^*(t_j)] = \mu_1(t_j)$ holds for ASPs, which is true at fully informative markers or if there are no disease genes linked to *t*. Misspecification of $\mathrm{Cov}(S_i^*|\Phi)$ reduces the asymptotic efficiency of the parameter estimates, but does not affect their consistency (Liang and Zeger 1986). An algorithm to solve this GEE is incorporated in a program called GENEFINDER (Liang et al. 2001*a*).

Chapter 3

The One-Locus Model: Modifications and Extensions of the Original GEE Method

For simplicity of presentation, in this chapter it will be assumed that the sample consists of independent ASPs; that is, there is only one ASP per nuclear family. In general, the discussion could easily be extended to multiple sib pairs per family.

Recall that in the one-locus method proposed by Liang et al. (2001*a*), estimates of $\delta_1 = (\tau, C)$ are obtained by solving the estimating equations:

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_1(\delta_1)}{\partial \delta_1} \right)' \mathrm{Cov}^{-1}(S_i^* \mid \Phi)(S_i^* - \mu_1(\delta_1)) = 0 \qquad (3.1)$$

where $S_i^* = (S_i^*(t_1), \ldots, S_i^*(t_M))' = (S_{i1}^*, \ldots, S_{iM}^*)'$ is the vector of estimated IBD sharing at all markers for the *i*th sib pair, given all marker data for the *i*th family.

Cov(S_i^*|Φ) can be written as $A_i^{1/2}R_iA_i^{1/2}$ where A_i is the $M \times M$ diagonal matrix of variances of the elements of the data vector for the *i*th sib pair S_i^*, expressed as a function of the expectation μ_{im}, and R_i is the $M \times M$ correlation matrix of S_i^*. If R_i is unknown or complicated, we may substitute it with $R_{0i}(\alpha)$, an $M \times M$ working correlation matrix for S_i^*, fully specified by a vector of unknown parameters, α. Although consistent estimates of the parameters and their standard errors can be obtained even if the choice of R_{0i} is not correct, careful modeling of Cov(S_i^*|Φ) may improve efficiency. In their formulation of the one-locus model GEE, Liang et al. (2001*a*) assumed an independence correlation structure and approximated the diagonal elements of Cov(S_i^*|Φ) by assuming that $\Pr(S_i(\tau)=0|\Phi) = 0$. In section 3.1 we consider approximating these diagonal elements of Cov(S_i^*|Φ) in different ways. In section 3.2 we introduce a different correlation structure that does not assume independence of marker IBD sharing at linked markers. In section 3.3 we extend the one-locus model of Liang et al. to estimate not only the excess mean ASP IBD sharing (*C*), but also the ASP identical-by-descent (IBD) sharing proportions at a single disease gene.

3.1 Alternative Formulations of Variances of Marker IBD Sharing

3.1.1 A simple variance approximation

In their implementation of the one-locus model, Liang et al. (2001*a*) used the independence working correlation, setting $R_{0i}(\alpha) = I$ (the identity matrix). This leads to a $M\times M$ diagonal matrix for Cov(S_i^*|Φ), with Var($S_i^*(t_m)$|Φ), m=1,…,M on the diagonal. Furthermore, in the original implementation Var($S_i^*(t_m)$|Φ) was approximated by Var($S_i(t_m)$|Φ), i.e. variances of fully informative marker IBD sharing. Liang et al. (2001*a*) showed that Var($S_i(t_m)|\Phi$) is a function of C, τ, and one additional parameter, p_0 = Pr($S(\tau)$=0|Φ):

$$\begin{aligned}\mathrm{Var}\left(S(t_m)\mid\Phi\right) &= \left(2\Psi-1\right)^2\left(\mathrm{Var}\left(S(\tau)\mid\Phi\right)-\frac{1}{2}\right)+\frac{1}{2}\\ &= e^{-.08|t_m-\tau|}\left(C-C^2+2p_0-\frac{1}{2}\right)+\frac{1}{2}\end{aligned} \qquad (3.2).$$

Note that under the null hypothesis, C = 0, p_0 = 0.25, and Var($S(\tau)|\Phi$) = Var($S(t_m)|\Phi$) = ½. In the original version of Genefinder, rather than estimating p_0, it was set to 0 and Var($S_i^*(t_m)|\Phi$) was approximated by:

$$\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m)\mid\Phi\right)=e^{-.08|t_m-\hat{\tau}|}\left(\hat{C}-\hat{C}^2-\frac{1}{2}\right)+\frac{1}{2} \qquad (3.3)$$

where $\hat{C}$ and $\hat{\tau}$ are the GEE estimates of the corresponding parameters.

Under some genetic models, the assumption p_0 = 0 may be quite accurate. However, under many models, it may lead to bias in estimates of Var($S_i^*(t_m)|\Phi$), especially at markers near the disease gene. For instance, if the penetrance vector is (0,0,1) and the frequency of the high-risk allele is 0.05, then (p_0,p_1,p_2) = (.002,.091,.907) and C = .905. In that case, Var($S(\tau)|\Phi$) = 0.090, and under the assumption that p_0 = 0, Var($S(\tau)|\Phi$) would be approximated to be 0.086. This approximation is quite accurate. However, if the penetrance vector is (0.001,0.001,0.2) with a disease allele frequency of 0.01, then (p_0,p_1,p_2) = (0.127,0.263,0.610), C = .483 and Var($S(\tau)|\Phi$) = 0.504, whereas the assumption that p_0 = 0 would lead to the approximation Var($S(\tau)|\Phi$) ≈ 0.250, which is not an accurate approximation.

3.1.2 Alternative specifications of Var($S^*(t_m)$|Φ): Corrected variance function and empirical variance estimation

Although it is known that the GEE approach is robust to misspecification of variances, we compared the performance of GEE estimation with the approximate variance formula shown in equation 3.3, to GEE estimation with other specifications of the variances of marker IBD sharing.

One alternative to the implementation in Genefinder is to again estimate Var($S_i^*(t_m)$|Φ) by Var($S_i(t_m)$|Φ), but to apply the exact formula for Var($S_i(t_m)$|Φ) (equation 3.2) by estimating the additional parameter p_0 = Pr($S(\tau)$=0|Φ) rather than setting it to zero. The parameter p_0 is not estimated by the solution of the GEE (equation 3.1) because the mean function, $\mu_1(t)$, does not directly depend on it. We estimate p_0 by the proportion of sib pairs in the sample sharing 0 alleles IBD at the marker nearest to τ. If this marker is not fully informative, we estimate the proportion of sib pairs sharing 0 alleles IBD by the proportion of sib pairs for which the estimated marker IBD sharing S_i^* is less than 0.5. That is

$$\hat{p}_0 \approx \hat{p}_{0k} = \hat{\Pr}(S(t_k) = 0 \mid \Phi) \approx \frac{\sum_{i=1}^{n} I(S_{ik}^* < 0.5)}{n} \qquad (3.4)$$

where t_k is the location of the kth marker, the marker closest to τ, and I is the indicator function. Using an estimate of p_0 (e.g. equation 3.4), Var($S_i^*(t_m)$|Φ) can be approximated by:

$$\hat{\mathrm{Var}}_{\mathrm{c}}\left(S(t_m) \mid \Phi\right) = e^{-.08|t_m-\hat{\tau}|}\left(\hat{C} - \hat{C}^2 + 2\hat{p}_0 - \frac{1}{2}\right) + \frac{1}{2} \ . \qquad (3.5)$$

Usually marker data is not fully informative, and the outcome variable for the ith ASP is the estimated IBD sharing, S_i^*, rather than the true IBD sharing, S_i, as is shown in equation 3.1. In both of the approaches described above the variance function for fully informative

IBD sharing, Var($S_i(t_m)$|Φ), is used to model variance of estimated IBD sharing, Var($S_i^*(t_m)$|Φ). To avoid this, Var($S_i^*(t_m)$|Φ) can be estimated empirically using data from *n* independent ASPs by:

$$\hat{\text{Var}}_{\text{emp}}(S_i^*(t_m) \mid \Phi) = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_i^*(t_m) - \bar{S}^*(t_m)\right)^2 \qquad (3.6)$$

where $\bar{S}^*(t_m) = \frac{\sum_{i=1}^{n} S_i^*(t_m)}{n}$,

or

$$\hat{\text{Var}}_{\text{emp}}(S_i^*(t_m) \mid \Phi) = \frac{1}{n-2}\sum_{i=1}^{n}\left(S_i^*(t_m) - \hat{\mu}(t_m)\right)^2 .$$

Benefits of this (ad-hoc) approach include: (1) the fact that variances of the observations do not need to be modeled, (2) computational simplicity, and (3) variances of the imputed data, S_i^*, are used rather than variances of the unknown marker IBD sharing, S_i. The main disadvantage is that these empirical variances may be less stable than modeled variances such as those in the first two approaches, and may therefore lead to larger sample size requirements. Using empirical variances of marker data is essentially equivalent to introducing a new variance parameter for each marker. Thus, if the number of markers is large, it may be more beneficial to model the variances rather than estimating them empirically.

Three FORTRAN programs were written for fitting the model for E($S(t)$|Φ) to marker IBD sharing data by solving the GEE (equation 3.1), implementing the three alternative formulations of Var($S_i^*(t_m)$|Φ).

1. GEE_GFv: Approximates Var($S_i^*(t_m)|\Phi$) by the simple variance formula implemented in the original version of GENEFINDER, $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m) \mid \Phi\right)$.
2. GEE_cv: Implements the "corrected" formula $\hat{\mathrm{Var}}_{\mathrm{c}}\left(S(t_m) \mid \Phi\right)$ to approximate Var($S_i^*(t_m)|\Phi$), by first estimating the additional parameter p_0.
3. GEE_empv: At marker t_m, Var($S_i^*(t_m)|\Phi$) is estimated using $\hat{\mathrm{Var}}_{\mathrm{emp}}(S_i^*(t_m) \mid \Phi)$, i.e. the sample variance of the estimated sharing among all the sib-pairs in the sample.

3.1.3 Simulations

Results of a small simulation study designed to compare estimates obtained by the one-locus GEE method with the three alternative variance formulations are shown in Tables 3.1 and 3.2. For each simulation, 1000 data sets were generated consisting of fully informative IBD sharing at either 11 or 21 equally spaced markers covering 100 cM. Samples of 100 or 500 ASPs were generated under two different genetic models, referred to as models A and B, with a single disease gene at $\tau = 50$ cM.

Although the simulations performed were not extensive, they did demonstrate that overall, in terms of bias and standard errors of estimates, and confidence interval coverage, GEEs with the three variance formulations had similar performance. For model B (Table 3.2), use of the simple variance function implemented in Genefinder, $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m) \mid \Phi\right)$, resulted in location estimates with smaller variances, and narrower confidence intervals. With the small sample size of 100 ASPs, for each of the three variance formulations, the average robust standard error of location estimates is smaller than the corresponding empirical standard deviation, indicating a downward bias of the robust standard errors. In large samples of 500 ASPs, however, an upward bias of the robust standard error estimates is observed. Most importantly, we can conclude that empirically estimating variances of S^* provides comparable results to those obtained by modeling the variance function, although for some underlying genetic models it may be less efficient.

Model	# ASPs	M	Var($S(t)$)	Estimation of C				Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
A	100	11	GFv	.477	.018	.068	.070	49.8	-0.2	3.21	2.75	.907
			Cv	.477	.018	.069	.071	49.8	-0.2	3.17	2.80	.905
			Empv	.479	.020	.069	.071	49.9	-0.1	3.06	2.76	.902
		21	GFv	.466	.007	.068	.068	50.0	0.0	2.77	2.40	.916
			Cv	.465	.006	.070	.069	50.0	0.0	2.90	2.51	.914
			Empv	.468	.009	.070	.068	50.0	0.0	2.76	2.45	.922
	500	11	GFv	.468	.009	.031	.031	49.9	-0.1	1.18	1.34	.919
			Cv	.468	.009	.032	.032	49.9	-0.1	1.20	1.36	.919
			Empv	.468	.009	.032	.031	49.9	-0.1	1.21	1.36	.920
		21	GFv	.464	.005	.031	.030	50.0	0.0	0.99	1.14	.949
			Cv	.464	.005	.032	.031	50.0	0.0	1.02	1.16	.948
			Empv	.465	.006	.032	.031	50.0	0.0	1.01	1.16	.947

Table 3.1: Comparison of different formulations of Var($S(t)$) in the one-locus GEE.

Model A: penetrance vector = (0.001,0.1,0.1), p(D) = 0.01, $C = 0.459$, ($p_0 = 0.021$), $\tau = 50.0$.

M=number of markers. Markers are equally spaced from 0 to 100 cM.

Var($S(t)$): GFv uses $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m)\mid\Phi\right)$, Cv uses $\hat{\mathrm{Var}}_{\mathrm{c}}\left(S(t_m)\mid\Phi\right)$, and Empv uses $\hat{\mathrm{Var}}_{\mathrm{emp}}(S_i^*(t_m)\mid\Phi)$ to estimate Var($S_i^*(t_m)|\Phi$).

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

Model	# ASPs	M	Var($S(t)$)	Estimation of C				Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
B	100	11	GFv	.497	.014	.083	.084	50.0	0.0	2.77	2.70	.914
			Cv	.494	.011	.087	.087	49.9	-0.1	3.07	2.95	.921
			Empv	.495	.012	.088	.087	50.0	0.0	3.07	2.94	.918
		21	GFv	.493	.010	.083	.082	49.9	-0.1	2.75	2.30	.905
			Cv	.492	.009	.087	.085	49.9	-0.1	3.05	2.64	.901
			Empv	.494	.011	.088	.085	49.9	-0.1	3.10	2.63	.897
	500	11	GFv	.495	.012	.038	.037	50.0	0.0	1.12	1.28	.928
			Cv	.493	.010	.040	.039	50.0	0.0	1.23	1.37	.931
			Empv	.493	.010	.040	.039	50.0	0.0	1.23	1.37	.929
		21	GFv	.487	.004	.036	.037	50.0	0.0	0.97	1.09	.946
			Cv	.486	.003	.037	.038	50.0	0.0	1.12	1.21	.945
			Empv	.486	.003	.038	.038	50.0	0.0	1.12	1.21	.945

Table 3.2: Comparison of different formulations of Var($S(t)$) in the one-locus GEE.

Model B: penetrance vector = (0.001,0.001,0.2), p(D) = 0.01, $C = 0.483$, ($p_0 = 0.127$), $\tau = 50.0$.

M=number of markers. Markers are equally spaced from 0 to 100 cM.

Var($S(t)$): GFv uses $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_j)\mid\Phi\right)$, Cv uses $\hat{\mathrm{Var}}_{\mathrm{c}}\left(S(t_j)\mid\Phi\right)$, and Empv uses $\hat{\mathrm{Var}}_{\mathrm{emp}}(S_i^*(t_m)\mid\Phi)$ to estimate Var($S_i^*(t_m)|\Phi$).

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

3.2 Modeling Correlation of IBD Sharing at Linked Markers

3.2.1 Introduction

For a pair of relatives, IBD sharing at two linked markers is correlated. Assuming fully informative markers and one disease gene in the region at position τ, Liang et al. (2001*a*) derived the following expressions for Cov($S(t_j)$,$S(t_l)$|Φ):

$$\mathrm{Cov}\left(S(t_j), S(t_l) \mid \Phi; H_A\right) =$$

$$\left(2\Psi_{t_l,t_j} - 1\right)\mathrm{Var}\left(S(\tau) \mid \Phi\right) \qquad \text{if } \tau \in \lfloor t_j, t_l \rfloor, \text{ and}$$

$$\left(2\Psi_{t_j,\tau} - 1\right)\left(2\Psi_{t_l,\tau} - 1\right)\left(\mathrm{Var}\left(S(\tau) \mid \Phi\right) - 1/2\right) + \frac{1}{2}\left(2\Psi_{t_j,t_l} - 1\right) \quad \text{if } \tau \notin \lfloor t_j, t_l \rfloor. \ (3.7)$$

Using Haldane's map function, the covariance can be re-written as:

$$\mathrm{Cov}\left(S(t_j), S(t_l) \mid \Phi; H_A\right) =$$

$$e^{-.04|t_j - t_l|}\left(C - C^2 + 2p_0\right) \qquad \text{if } \tau \in \lfloor t_j, t_l \rfloor, \text{ and}$$

$$e^{-.04|t_j - \tau| - .04|t_l - \tau|}\left(C - C^2 + 2p_0 - \frac{1}{2}\right) + \frac{1}{2}e^{-.04|t_j - t_l|} \qquad \text{if } \tau \notin \lfloor t_j, t_l \rfloor.$$

Under the null hypothesis of no linkage (i.e. $\tau = \infty$), the above expression reduces to:

$$\mathrm{Cov}\left(S(t_j), S(t_l) \mid \Phi; H_0\right) = \frac{1}{2}\left(2\Psi_{t_l,t_j} - 1\right) = \frac{1}{2}e^{-.04|t_j - t_l|}.$$

Despite the fact that IBD sharing status at linked markers is known to be correlated, in the GENEFINDER software program (Liang et al. 2001*a*), the independence working correlation structure was used. With this working correlation Cov(S_i|Φ) = $\mathrm{A}_i^{1/2}\mathrm{R}_\mathrm{I}\mathrm{A}_\mathrm{i}^{1/2}$ where R_I = I is the in the M×M identity matrix. This model simplifies computations considerably. Even if the correlation is misspecified, consistent estimates of the parameters and their variances can be obtained by solving generalized estimating equations and using the robust “sandwich” estimator for variance estimation (Liang and Zeger 1986). However, efficiency of the estimators may be reduced by assuming an incorrect correlation structure for the observations. We investigated the potential improvement in the efficiency of estimates by incorporating correlation between data at linked markers.

3.2.2 Alternative correlation structures

Although the exact correlation of IBD sharing for a pair of affected siblings at two fully informative linked markers is known (equations 3.7), we consider an approximate correlation structure that simplifies computation. The justification is that: first, even if the correlation is misspecified, estimates will remain consistent; and second, incorporating a new correlation structure, even if not exactly correct, may improve properties of the estimates if that correlation structure more closely approximates the true correlation structure than the assumption of complete independence. One possible correlation structure is the AR-1 model:

$$R_{AR-1} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{m-1} \\ \rho & 1 & \rho & \cdots & \rho^{m-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \rho^{m-3} & \cdots & 1 \end{bmatrix}$$

This correlation model assumes that the correlation of IBD sharing at consecutive markers is ρ, at markers separated by one marker is ρ^2, etc. Because it assumes equal correlation for any two consecutive markers, this model is more appropriate for equally spaced markers. Also, it ignores the fact that the level of correlation depends on the positions of the markers relative to the position of the disease gene. However, $R_{AR\text{-}1}$ seems to be a reasonable approximation to the true correlation matrix, because it specifies that the correlation of IBD sharing at two particular markers decreases as a function of the number of markers between those two markers. This means that correlation decreases as a function of distance between markers, if they are equally spaced. Other useful features of $R_{AR\text{-}1}$ include the fact that it has a simple inverse, leading to simpler computations. Furthermore, because the correlation does not depend on the position of the disease gene, the covariance matrix only needs to be computed once.

The inverse of the correlation matrix $R_{AR\text{-}1}$ is

$$R_{AR-1}{}^{-1} = \begin{bmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & \vdots \\ 0 & -\rho & \ddots & -\rho & \vdots \\ \vdots & \vdots & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots & 0 & -\rho & 1 \end{bmatrix}$$

That is,

$$R_{AR-1}{}^{-1}(i,j) = \begin{cases} 1 & \text{for } i=j=1 \text{ and } i=j=m \\ 1+\rho^2 & \text{for } i=j=2,\dots,m-1 \\ -\rho & \text{for } |i-j|=1 \\ 0 & \text{otherwise} \end{cases}$$

With this correlation matrix, $\operatorname{Cov}\left(S_i\right) = A^{1/2} R_{AR-1} A^{1/2}$

where A is the diagonal matrix of variances of IBD sharing at the markers:

$$A = \begin{bmatrix} \operatorname{Var}(S(t_1)\mid\Phi) & 0 & 0 & \dots & 0 \\ 0 & \operatorname{Var}(S(t_2)\mid\Phi) & 0 & \dots & 0 \\ 0 & 0 & \operatorname{Var}(S(t_3)\mid\Phi) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \operatorname{Var}(S(t_m)\mid\Phi) \end{bmatrix}.$$

Thus

$$\operatorname{Cov}^{-1}\left(S_i\right) = \left(A^{1/2}\right)^{-1} R_{AR-1}{}^{-1} \left(A^{1/2}\right)^{-1}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{v_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{v_2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{v_M}} \end{bmatrix} \begin{bmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & \vdots \\ 0 & -\rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots & 0 & -\rho & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{v_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{v_2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{v_M}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{v_1} & \frac{-\rho}{\sqrt{v_1 v_2}} & 0 & \dots & 0 \\ \frac{-\rho}{\sqrt{v_1 v_2}} & \frac{1+\rho^2}{v_2} & \frac{-\rho}{\sqrt{v_2 v_3}} & \dots & \vdots \\ 0 & \frac{-\rho}{\sqrt{v_2 v_3}} & \ddots & \frac{-\rho}{\sqrt{v_{M-2} v_{M-1}}} & \vdots \\ \vdots & \vdots & \frac{-\rho}{\sqrt{v_{M-2} v_{M-1}}} & \frac{1+\rho^2}{v_{m-1}} & \frac{-\rho}{\sqrt{v_{M-1} v_M}} \\ 0 & \dots & 0 & \frac{-\rho}{\sqrt{v_{M-1} v_M}} & \frac{1}{v_M} \end{bmatrix}$$

where $v_i = \mathrm{Var}(S(t_i)|\Phi)$.

That is,

$$\mathrm{Cov}^{-1}(S(t_i), S(t_j)) = \begin{cases} \frac{1}{v_i} & \text{for } i = j = 1 \text{ and } i = j = m \\ \frac{1+\rho^2}{v_i} & \text{for } i = j = 2, \dots, m-1 \\ \frac{-\rho}{\sqrt{v_i v_j}} & \text{for } |i - j| = 1 \\ 0 & \text{otherwise} \end{cases}$$

This working covariance model was implemented in a new FORTRAN program for solving the one-locus model GEE.

3.2.3 Simulations

Results of simulations designed to compare estimation of the location and expected IBD sharing of a disease gene by the GEE method, with two different working correlations for the data, are shown in Tables 3.3 and 3.4 (using the simplified variance function $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m) \mid \Phi\right)$) and Tables 3.5 and 3.6 (using empirical variances $\hat{\mathrm{Var}}_{\mathrm{emp}}(S_i^*(t_m) \mid \Phi)$). For each simulation, 1000 replicate data sets were generated. The same genetic models, marker maps, and sample sizes were used as for the simulations presented in section 3.1.3.

Standard deviations for the estimates using the two approaches reveal that using the AR-1 working correlation rather than the independence working correlation led to improved efficiency of estimators. Although estimates of mean excess IBD sharing at the disease locus (*C*) were more biased when the AR-1 working correlation structure was applied, the mean squared error tended to be lower. For most of the simulations performed, the independence correlation model led to location confidence interval coverage slightly closer to the nominal 95%, than the AR-1 model. However, the AR-1 model resulted in closer-to-nominal CI coverage in large samples with more genotyped markers (500 ASPs, and 21 markers spanning 100cM). Thus, it appears that modeling the correlation of IBD sharing data at linked markers, even with a simple approximation such as the AR-1 model, has the potential to improve estimation of disease gene locations, by providing narrower confidence intervals with closer to nominal coverage. However, it is also apparent that larger samples are required to apply this more complex covariance model.

Model	# ASPs	M	Corr(S_i)	Estimation of C				Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
A	100	11	I	.472	.013	.067	.070	50.0	0.0	3.20	2.80	0.908
			AR-1	.490	.029				-0.1	2.07	2.16	0.905
		21	I	.469	.010	.067	.067	50.0	0.0	2.58	2.40	0.927
			AR-1	.489	.028				-0.1	1.81	1.42	0.908
	500	11	I	.468	.009	.031	.031	50.1	0.1	1.17	1.34	0.941
			AR-1	.478	.019				0.1	0.89	1.07	0.927
		21	I	.462	.003	.030	.030	50.0	0.0	1.06	1.14	0.932
			AR-1	.474	.015				0.0	0.56	0.73	0.940

Table 3.3: Comparison of two different working correlations in the one-locus GEE with a simple variance approximation. I = independence model; AR-1 = first order autoregressive model. For these simulations the simplified variance formula $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_j)\,|\,\Phi\right)$ was used. M=number of markers. Markers are equally spaced from 0 to 100 cM.

Model A: penetrance vector = (0.001,0.1,0.1), p(D) = 0.01, $C = 0.459$, $\tau = 50.0$.

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

Model	# ASPs	M	Corr(S_i)	Estimation of C				Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
B	100	11	I	.498	.015	.085	.084	49.9	-0.1	2.96	2.70	0.919
			AR-1	.518	.035	.076	.078	50.0	0.0	2.10	2.06	0.906
		21	I	.492	.009	.079	.081	50.1	0.1	2.60	2.32	0.915
			AR-1	.512	.029	.073	.075	49.9	-0.1	1.48	1.37	0.908
	500	11	I	.492	.009	.037	.037	50.0	0.0	1.11	1.28	0.941
			AR-1	.503	.020	.033	.035	50.0	0.0	0.85	1.03	0.926
		21	I	.485	.002	.037	.037	50.0	0.0	1.01	1.09	0.933
			AR-1	.499	.016	.033	.033	50.0	0.0	0.53	0.70	0.944

Table 3.4: Comparison of two different working correlations in the one-locus GEE with a simple variance approximation. I = independence model; AR-1 = first order autoregressive model. For these simulations the simplified variance formula $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_j)\,|\,\Phi\right)$ was used. M=number of markers. Markers are equally spaced from 0 to 100 cM.

Model B: penetrance vector = (0.001,0.001,0.2), p(D) = 0.01, C = 0.483, τ = 50.0.

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

Model	# ASPs	M	Corr(S_i)	Estimation of C				Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
A	100	11	I	.474	.015	.068	.071	49.9	-0.1	3.03	2.82	0.904
			AR-1	.491	.032	.056	.062	49.9	-0.1	2.16	2.16	0.899
		21	I	.471	.012	.068	.068	50.0	0.0	2.57	2.44	0.933
			AR-1	.495	.036	.057	.057	50.0	0.0	1.47	1.39	0.904
	500	11	I	.468	.009	.031	.031	50.1	0.1	1.19	1.36	0.941
			AR-1	.478	.019	.026	.027	50.1	0.1	0.89	1.07	0.926
		21	I	.463	.004	.030	.031	50.0	0.0	1.09	1.16	0.931
			AR-1	.475	.016	.025	.025	50.0	0.0	0.57	0.73	0.938

Table 3.5: Comparison of two different working correlations in the one-locus GEE with empirical variance approximation. I = independence model; AR-1 = first order autoregressive model. For these simulations empirical variances $\hat{\mathrm{Var}}_{\mathrm{emp}}(S_i^*(t_m)\,|\,\Phi)$ were used.

M=number of markers. Markers are equally spaced from 0 to 100 cM.

Model A: penetrance vector = (0.001,0.1,0.1), p(D) = 0.01, $C = 0.459$, $\tau = 50.0$.

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

Model	# ASPs	M	Corr(S_i)	Estimation of C				Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
B	100	11	I	.496	.013	.090	.087	49.9	-0.1	3.24	2.94	0.915
			AR-1	.513	.032	.078	.076	50.0	0.0	2.16	2.09	0.904
		21	I	.492	.009	.084	.085	50.1	0.1	2.97	2.64	0.904
			AR-1	.512	.029	.074	.072	50.0	0.0	1.41	1.36	0.896
	500	11	I	.490	.007	.039	.039	50.0	0.0	1.22	1.37	0.938
			AR-1	.499	.016	.033	.034	50.0	0.0	0.84	1.04	0.927
		21	I	.484	.001	.038	.038	50.0	0.0	1.17	1.22	0.931
			AR-1	.495	.012	.033	.033	50.0	0.0	0.53	0.71	0.941

Table 3.6: Comparison of two different working correlations in the one-locus GEE with empirical variance approximation. I = independence model; AR-1 = first order autoregressive model. For these simulations empirical variances $\hat{\text{Var}}_{\text{emp}}(S_i^*(t_m)\,|\,\Phi)$ were used.

M=number of markers. Markers are equally spaced from 0 to 100 cM.

Model B: penetrance vector = (0.001,0.001,0.2), p(D) = 0.01, $C = 0.483$, $\tau = 50.0$.

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

3.3 Estimation of a Disease Gene Location and IBD Sharing Probabilities for ASPs

3.3.1 Introduction

The methods described in section 2.3 and extended in sections 3.1 and 3.2 may lead to some loss of information for particular underlying genetic models. When markers are not fully informative, multipoint marker data can be used to estimate the statistics z_0, z_1, and z_2 for each ASP at each marker, where z_j is the estimated (posterior) probability that the pair shares j alleles IBD. Such calculations can be carried out, for example, using the software GENEHUNTER. However $C = \mathrm{E}[S(\tau)|\Phi]\text{-}1$ and τ are estimated using only the (estimated) IBD sharing (S^*) for each ASP at each marker. That is, the methods proposed by Liang et al. (2001*a*) do not use the z_j's, but rather a linear combination of them, $S^* = z_1+2*z_2$, as the outcome for each sib pair at each marker. Here, we propose treating $(z_1,z_2)_{im}$ as the outcome at marker *m* for sib pair *i*. This method allows the estimation of not only C and τ, but also the IBD sharing probabilities (p_0,p_1,p_2) where p_j=Pr($S(\tau)$=$j|\Phi$).

3.3.2 Probabilities of sharing 0, 1, and 2 alleles IBD by ASPs at linked loci

Let $\pi_0(t)$, $\pi_1(t)$, and $\pi_2(t)$ denote the probabilities that an ASP shares 0, 1, or 2, alleles IBD at a locus t. Under the null hypothesis of no linkage, $(\pi_0(t),\pi_1(t),\pi_2(t)) = (1/4,1/2,1/4)$. If t is linked to a disease gene, there is a distortion from these null probabilities. Assume that there is a disease gene at τ, and let p_0, p_1, and p_2 denote the probabilities that an ASP shares 0, 1, or 2 alleles IBD at τ, i.e. $p_j = \Pr(S(\tau)=j|\Phi) = \pi_j(\tau)$.

The values ($\pi_0(t),\pi_1(t),\pi_2(t)$) have a simple dependence on (p_0, p_1, p_2) and the distance between t and τ. Let $\Psi = \Psi_{t,\tau} = \theta_{t,\tau}^2 + (1-\theta_{t,\tau})^2$ be the measure of genetic map distance between t and τ. The relationship between ($\pi_0(t),\pi_1(t),\pi_2(t)$) and ($p_0,p_1,p_2$) can be derived as follows:

$$\pi_j(t) = \Pr\left(S(t) = j \mid \Phi\right) = \sum_{l=0}^{2} \Pr\left(S(t) = j, S(\tau) = l \mid \Phi\right)$$

$$= \sum_{l=0}^{2} \Pr\left(S(t) = j \mid S(\tau) = l, \Phi\right) \Pr\left(S(\tau) = l \mid \Phi\right)$$

$$= \sum_{l=0}^{2} \Pr\left(S(t) = j \mid S(\tau) = l\right) \Pr\left(S(\tau) = l \mid \Phi\right)$$

The conditional IBD distribution at any locus t, given IBD status at a linked locus τ, was derived by Haseman and Elston (1972) and is shown in Table 3.7. Substituting expressions from Table 3.7 for the conditional probabilities leads to:

$$\pi_1(t) = \Pr\left(S(t) = 1 \mid \Phi\right) = \sum_{l=0}^{2} \Pr\left(S(t) = 1 \mid S(\tau) = l\right) \Pr\left(S(\tau) = l \mid \Phi\right)$$

$$= 2\Psi(1-\Psi)p_0 + \left(1 - 2\Psi(1-\Psi)\right)p_1 + 2\Psi(1-\Psi)p_2$$

$$= 2\Psi(1-\Psi)(1 - p_1 - p_2) + \left(1 - 2\Psi(1-\Psi)\right)p_1 + 2\Psi(1-\Psi)p_2$$

$$= 2\Psi(1-\Psi) + \left(1 - 4\Psi(1-\Psi)\right)p_1$$

and

	IBD$_t$= 0	IBD$_t$= 1	IBD$_t$= 2
IBD$_\tau$= 0	Ψ^2	2Ψ(1-Ψ)	(1-Ψ)^2
IBD$_\tau$= 1	Ψ(1-Ψ)	1-2Ψ(1-Ψ)	Ψ(1-Ψ)
IBD$_\tau$= 2	(1-Ψ)^2	2Ψ(1-Ψ)	Ψ^2

Table 3.7 Conditional IBD distribution at t given IBD status at τ: Pr(IBD$_t$=i| IBD$_\tau$=j)

$$\pi_2(t) = \Pr\big(S(t)=2 \mid \Phi\big) = \sum_{l=0}^{2} \Pr\big(S(t)=2 \mid S(\tau)=l\big)\Pr\big(S(\tau)=l \mid \Phi\big)$$

$$= (1-\Psi)^2 p_0 + \Psi(1-\Psi)p_1 + \Psi^2 p_2$$

$$= (1-\Psi)^2(1-p_1-p_2) + \Psi(1-\Psi)p_1 + \Psi^2 p_2$$

$$= (1-\Psi)^2 + (1-\Psi)(2\Psi-1)p_1 + (2\Psi-1)p_2$$

The formula for $\pi_0(t)$ is omitted since $\pi_0(t) + \pi_1(t) + \pi_2(t) = 1$. Haldane's map function can be used to convert recombination fractions to map distance ($\theta_{t_1,t_2} = \big(1-e^{-0.02|t_1-t_2|}\big)/2$), leading to the following expressions for $\pi_1(t)$ and $\pi_2(t)$:

$$\pi_1(t) = \frac{1}{2} + e^{-.08|t-\tau|}\left(p_1 - \frac{1}{2}\right) \text{ and}$$

$$\pi_2(t) = \frac{1}{4} + e^{-.04|t-\tau|}\left(p_2 + \frac{1}{2}p_1 - \frac{1}{2}\right) + e^{-.08|t-\tau|}\left(\frac{1}{4} - \frac{1}{2}p_1\right).$$

Note that if t is linked to location τ, but τ is not linked to a disease gene and thus has null IBD sharing proportions, then p_1=1/2 and p_2=1/4, and this leads to π_1(t)=1/2 and π_2(t)=1/4. Also, if τ is a disease gene, but t is not linked to τ or any other disease gene, then $|t-\tau| = \infty$ which again leads to $\pi_1(t)$=1/2 and $\pi_2(t)$=1/4. This verifies that for a locus t not linked to a disease gene $(\pi_0(t),\pi_1(t),\pi_2(t))$ = (1/4,1/2,1/4) as expected.

Plots of $\pi_1(t)$ and $\pi_2(t)$ curves for two models are shown in Figure 3.1. Note that for any t, $\pi_1(t) + 2\pi_2(t) = \mu_1(t)$ where $\mu_1(t)$=E[$S(\tau)|\Phi$] under the one-locus model.

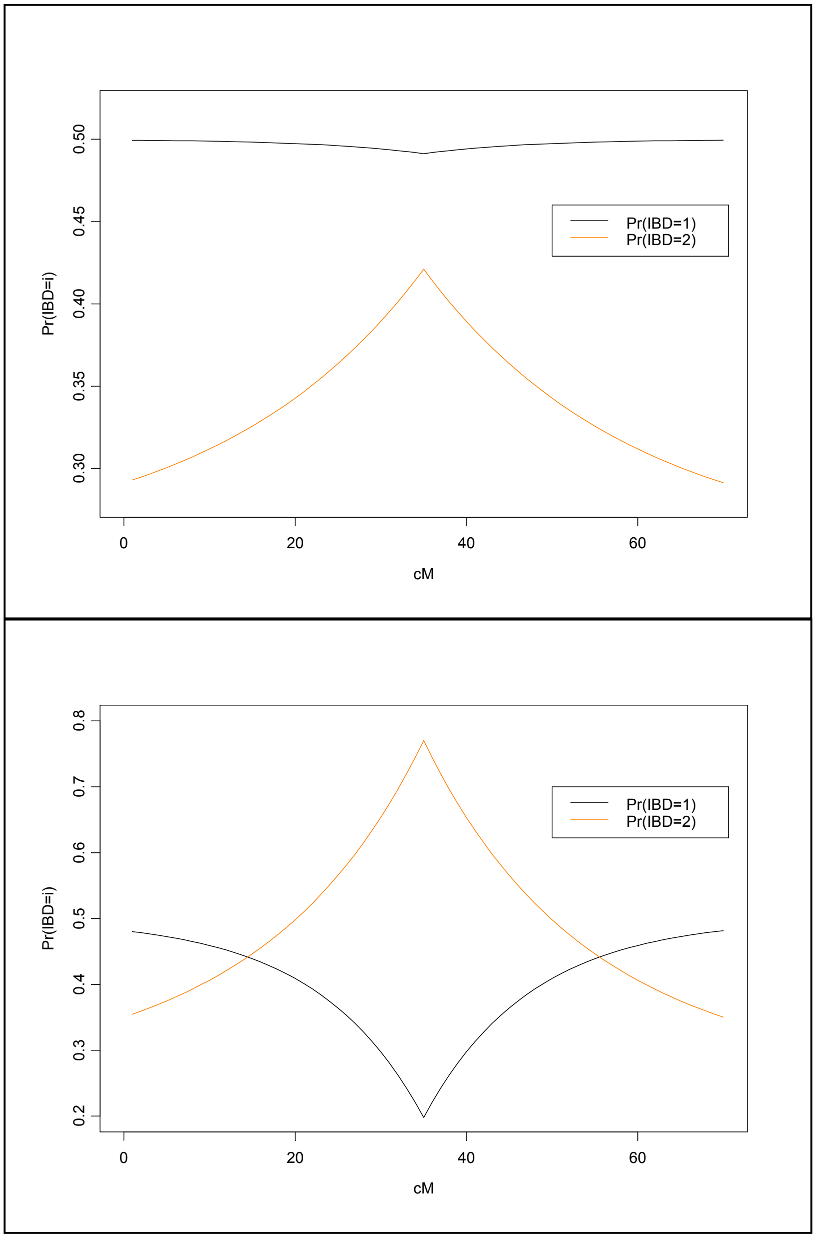


Figure 3.1: IBD sharing probability curves across a chromosome containing a disease gene.

a) Model: $\tau = 35.0$, penetrances = (.01,.9,.9), p(D) = 0.1, $p_1 = 0.491228$, $p_2 = 0.421059$

b) Model: $\tau = 35.0$, penetrances = (.01,.01,.95), p(D) = 0.1, $p_1 = 0.197846$, $p_2 = 0.770376$

(D represents the disease predisposing allele)

3.3.3 Distribution, expectation and variance of the IBD sharing statistics

Let $(z_0(t_m),z_1(t_m),z_2(t_m))_i$ denote the posterior probabilities of sharing 0, 1, and 2 alleles IBD at position t_m (marker m) by the ith ASP, given all the marker information for family i. If marker m is fully informative for the ith ASP, then each $z_j(t_m)$ equals 0 or 1 such that $z_0 + z_1 + z_2 = 1$; i.e. $(z_0(t_m),z_1(t_m),z_2(t_m))_i = (1,0,0)$ or $(0,1,0)$ or $(0,0,1)$. In that case, $(z_0(t_m),z_1(t_m),z_2(t_m))_i$ is multinomial with $n = 1$ and probabilities $(\pi_0(t_m),\pi_1(t_m),\pi_2(t_m))$. Because the ASPs are independent, $(z_0(t_m),z_1(t_m),z_2(t_m))_i$ is independent of $(z_0(t_m),z_1(t_m),z_2(t_m))_j$ for $i\neq j$. However, $(z_0(t_m),z_1(t_m),z_2(t_m))_i$ is not independent of $(z_0(t_l),z_1(t_l),z_2(t_l))_i$ for $m\neq l$ if markers m and l are linked.

Now note that at a fully informative marker, or if there are no disease genes linked to t, then $E[z_j(t)|\Phi] = \pi_j(t)$. Therefore, we may estimate $\pi_j(t)$ by $z_j(t)$. Also, under the assumption of fully informative markers, because $(z_0(t_m),z_1(t_m),z_2(t_m))_i$ is multinomial with n=1 and probabilities $(\pi_0(t_m),\pi_1(t_m),\pi_2(t_m))$, at any locus for any sib-pair:

$\mathrm{Var}[z_j(t)] = \pi_j(t)(1-\pi_j(t))$ and

$\mathrm{Cov}[z_j(t),z_k(t)] = -\pi_j(t)\pi_k(t)$.

3.3.4 Estimation of the disease gene location and IBD sharing probabilities at the disease gene

It was shown that $\pi_j(t)$, which is a function of the parameters (p_0,p_1,p_2) and τ, can be estimated by the observed (or estimated) $z_j(t)$. Since $z_0+z_1+z_2=1$, (z_0,z_1,z_2) can be fully specified by (z_1,z_2). Similarly, only two of the p_j parameters are required to fully specify the IBD sharing probabilities at τ. Let $\eta = (p_1,p_2,\tau)$. Then $(\pi_1(t), \pi_2(t))' = E[z_1(t), z_2(t)]'$, is a function of η. Because of the correlation between $z_1(t)$ and $z_2(t)$ for any sib pair at a given marker, and the dependence of observations across markers within a family, the generalized estimating equations approach, which accounts for dependent observations, is an appealing choice of analysis method to apply for estimation of η. Note that since $C = p_1 + 2p_2 - 1$, estimation of $\eta = (p_1, p_2, \tau)$ also provides an estimate of mean excess IBD sharing, C. However, estimation of $\delta = (C, \tau)$ does not provide estimates of the IBD sharing probabilities (p_0, p_1, p_2).

Now let $z_i = (z_{1i1}, z_{2i1}, z_{1i2}, z_{2i2}, \ldots, z_{1iM}, z_{2iM})$' where z_{jim} is the estimated (posterior) probability of sharing j alleles IBD by sibpair i at marker m (position t_m). Estimates of $\eta = (p_1, p_2, \tau)$ can be obtained by solving the following estimating equations:

$$\sum_{i=1}^{n} \left(\frac{\partial \pi(\eta)}{\partial \eta} \right)' \text{Cov}^{-1}(z_i \mid \Phi)(z_i - \pi(\eta)) = 0 \tag{3.8}$$

where $\pi(\eta) = \text{E}[z_i|\Phi] = (\pi_1(t_1), \pi_2(t_1), \pi_1(t_2), \pi_2(t_2), \ldots, \pi_1(t_M), \pi_2(t_M))'$. Derivatives of the mean functions $\pi_1(t)$ and $\pi_2(t)$ with respect to the parameters are:

$$\frac{\partial \pi_1(t)}{\partial p_1} = e^{-.08|t-\tau|}$$

$$\frac{\partial \pi_1(t)}{\partial p_2} = 0$$

$$\frac{\partial \pi_1(t)}{\partial \tau} = \text{sgn}(.08)\left(p_1 - \frac{1}{2} \right) e^{-.08|t-\tau|} \quad \text{where} \quad \text{sgn} = \begin{cases} +1 & t > \tau \\ -1 & t < \tau \end{cases}$$

$$\frac{\partial \pi_2(t)}{\partial p_1} = \frac{e^{-.04|t-\tau|}}{2} - \frac{e^{-.08|t-\tau|}}{2}$$

$$\frac{\partial \pi_2(t)}{\partial p_2} = e^{-.04|t-\tau|}$$

$$\frac{\partial \pi_2(t)}{\partial \tau} = \text{sgn}(.04)\left(p_2 + \frac{1}{2} p_1 - \frac{1}{2} \right) e^{-.04|t-\tau|} + \text{sgn}(.08)\left(\frac{1}{4} - \frac{1}{2} p_1 \right) e^{-.08|t-\tau|}$$

$$\text{where sgn} = \begin{cases} +1 & t > \tau \\ -1 & t < \tau \end{cases}$$

If $z_i = (z_{1i1}, z_{2i1}, z_{1i2}, z_{2i2}, \ldots, z_{1iM}, z_{2iM})$' denotes all the observed/estimated IBD sharing probabilities for the *i*th sib pair, then z_i is independent of z_j. However, (z_{1im}, z_{2im}) is not independent of (z_{1il}, z_{2il}) since IBD sharing at linked markers is correlated. Also, the two components of the vector (z_{1im}, z_{2im}) are not independent.

Since the GEE approach provides consistent estimates of the parameters even if the correlation structure for the observations z_i is misspecified (Liang and Zeger 1986), the simplest correlation matrix that could be used would be the identity correlation matrix. This correlation model ignores all dependencies in the data. Although this simple working correlation would provide consistent estimates of the parameters, incorrect specification of the correlation structure may lead to decreased efficiency. One simple adjustment of the correlation structure takes into account the correlation between z_{1im} and z_{2im} for each marker (m) and sib pair (i).

It follows from the discussion in section 3.3.3, that under the assumption of fully informative markers, $\mathrm{Cov}[z_{1im}, z_{2im}] = -\pi_1(t_m)\pi_2(t_m)$. Taking this covariance into account, leads to a block diagonal $\mathrm{Cov}(z_i)$, with 2×2 blocks on the diagonal as follows:

$$\mathrm{Cov}(z_i \mid \Phi) =$$

$$= \begin{bmatrix} \mathrm{Var}(z_{1i1}) & \mathrm{Cov}(z_{1i1}, z_{2i1}) & 0 & 0 & \cdots & 0 & 0 \\ \mathrm{Cov}(z_{1i1}, z_{2i1}) & \mathrm{Var}(z_{2i1}) & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \mathrm{Var}(z_{1i2}) & \mathrm{Cov}(z_{1i2}, z_{2i2}) & \cdots & 0 & 0 \\ 0 & 0 & \mathrm{Cov}(z_{1i2}, z_{2i2}) & \mathrm{Var}(z_{2i2}) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mathrm{Var}(z_{1iM}) & \mathrm{Cov}(z_{1iM}, z_{2iM}) \\ 0 & 0 & 0 & 0 & \cdots & \mathrm{Cov}(z_{1iM}, z_{2iM}) & \mathrm{Var}(z_{2iM}) \end{bmatrix}$$

$$= \begin{bmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{21} & 0 & 0 & \cdots & 0 & 0 \\ -\pi_{11}\pi_{21} & \pi_{21}(1-\pi_{21}) & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{22} & \cdots & 0 & 0 \\ 0 & 0 & -\pi_{12}\pi_{22} & \pi_{22}(1-\pi_{22}) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \pi_{1M}(1-\pi_{1M}) & -\pi_{1M}\pi_{2M} \\ 0 & 0 & 0 & 0 & \cdots & -\pi_{1M}\pi_{2M} & \pi_{2M}(1-\pi_{2M}) \end{bmatrix}$$

where $\pi_{jm} = \pi_j(t_m)$.

3.3.5 Implementation and simulations

An algorithm to solve the GEE for estimating η (equation 3.8) was implemented in a FORTRAN program, and simulations were carried out to study the performance of this estimation approach using the same models, samples sizes, and marker maps as in sections 3.1.3 and 3.2.3. Results based on 1000 replicates are shown in Tables 3.8-3.11.

One advantage of the new approach is that estimation of $\eta = (p_1, p_2, \tau)$ also indirectly provides estimates of mean excess IBD sharing, C, whereas estimation of $\delta = (C, \tau)$ does not provide estimates of the IBD sharing proportions (p_0, p_1, p_2). Although estimates of C tend to be biased up, this bias decreases when the number of markers or number of ASPs in the sample are increased. Also, estimates of C are less biased when they are calculated from the IBD sharing proportion estimates (η-parameterization). In large samples (500 ASPs), estimation of $\eta = (p_1,p_2,\tau)$ led to more precise estimates of τ than estimation of $\delta = (C,\tau)$, as is demonstrated by the standard deviation estimates in Tables 3.10 and 3.11. Using the GEE approach to estimate $\eta = (p_1,p_2,\tau)$ led to greater confidence interval under-coverage for the gene location (τ) than estimation of $\delta = (C,\tau)$. However, the difference in CI coverage diminished as the sample size was increased.

Model	# ASPs	M	Parameters	Estimation of C				Estimation of p_1				Estimation of p_2			
				Average estimate	Bias	SD_{emp}	SE_{rob}	Average estimate	Bias	SD_{emp}	SE_{rob}	Average estimate	Bias	SD_{emp}	SE_{rob}
A	100	11	(C,τ)	.477	.018	.068	.070	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.468	.009	.106	--	.496	-.003	.073	.063	.486	.006	.068	.063
		21	(C,τ)	.466	.007	.068	.068	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.463	.004	.069	--	.499	.000	.066	.063	.482	.002	.063	.062
	500	11	(C,τ)	.468	.009	.031	.031	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.466	.007	.031	--	.503	.004	.026	.027	.482	.002	.027	.028
		21	(C,τ)	.464	.005	.031	.030	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.463	.004	.032	--	.499	.000	.028	.028	.482	.002	.029	.028

Table 3.8: Properties of estimates of C, p_1, and p_2 from GEE estimation of (C,τ) and GEE estimation of (p_1,p_2,τ). The simplified variance formula $\hat{\text{Var}}_{\text{GF}}\left(S(t_j)\,|\,\Phi\right)$ was used in estimation of (C,τ) and multinomial variances (see text) of z_1 and z_2 were used in estimation of (p_1,p_2,τ).

M=number of markers. Markers are equally spaced from 0 to 100 cM.

SD_{emp} = empirical standard deviation estimate. SE_{rob} = average robust standard error estimate.

Model A: penetrance vector = (0.001,0.1,0.1), p(D) = 0.01, $C = 0.459$, ($p_0 = 0.021$, $p_1 = 0.499$, $p_2 = 0.480$), $\tau = 50.0$.

Note: For (p_1,p_2,τ)-parameterization, C was estimated by: $\hat{C} = \hat{p}_1 + 2\hat{p}_2 - 1$.

Model	# ASPs	M	Parameters	Estimation of C				Estimation of p_1				Estimation of p_2			
				Average estimate	Bias	SD_{emp}	SE_{rob}	Average estimate	Bias	SD_{emp}	SE_{rob}	Average estimate	Bias	SD_{emp}	SE_{rob}
B	100	11	(C,τ)	.497	.014	.083	.084	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.489	.006	.087	--	.247	-.016	.059	.063	.621	.011	.062	.065
		21	(C,τ)	.493	.010	.083	.082	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.490	.007	.087	--	.253	-.010	.057	.057	.618	.008	.061	.061
	500	11	(C,τ)	.495	.012	.038	.037	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.491	.008	.040	--	.254	-.009	.026	.028	.618	.008	.029	.030
		21	(C,τ)	.487	.004	.036	.037	--	--	--	--	--	--	--	--
			(p_1,p_2,τ)	.485	.002	.038	--	.260	-.003	.025	.026	.613	.003	.026	.028

Table 3.9: Properties of estimates of C, p_1, and p_2 from GEE estimation of (C,τ) and GEE estimation of (p_1,p_2,τ). The simplified variance formula $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_j)\,|\,\Phi\right)$ was used in estimation of (C,τ) and multinomial variances (see text) of z_1 and z_2 were used in estimation of (p_1,p_2,τ).

M=number of markers. Markers are equally spaced from 0 to 100 cM.

SD_{emp} = empirical standard deviation estimate. SE_{rob} = average robust standard error estimate.

Model B: penetrance vector = (0.001,0.001,0.2), p(D) = 0.01, $C = 0.483$, ($p_0 = 0.127$, $p_1 = 0.263$, $p_2 = 0.610$), $\tau = 50.0$.

Note: For (p_1,p_2,τ)-parameterization, C was estimated by: $\hat{C} = \hat{p}_1 + 2\hat{p}_2 - 1$.

Model	# ASPs	M	Parameters	Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
A	100	11	(C,τ)	49.8	-0.2	3.21	2.75	0.907
			(p_1,p_2,τ)	49.6	-0.4	5.28	2.15	0.818
		21	(C,τ)	50.0	0.0	2.77	2.40	0.916
			(p_1,p_2,τ)	50.0	0.0	2.80	1.71	0.816
	500	11	(C,τ)	49.9	-0.1	1.18	1.34	0.919
			(p_1,p_2,τ)	49.9	-0.1	1.00	1.03	0.914
		21	(C,τ)	50.0	0.0	0.99	1.14	0.949
			(p_1,p_2,τ)	50.0	0.0	0.82	0.80	0.939

Table 3.10: Properties of estimates of τ from GEE estimation of (C,τ) and GEE estimation of (p_1,p_2,τ). The simplified variance formula $\hat{\text{Var}}_{\text{GF}}\left(S(t_j)\,|\,\Phi\right)$ was used in estimation of (C,τ) and multinomial variances (see text) of z_1 and z_2 were used in estimation of (p_1,p_2,τ).

M=number of markers. Markers are equally spaced from 0 to 100 cM.

Model A: penetrance vector = (0.001,0.1,0.1), p(D) = 0.01, $C = 0.459$, ($p_0 = 0.021$, $p_1 = 0.499$, $p_2 = 0.480$), $\tau = 50.0$.

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

Model	# ASPs	M	Parameters	Estimation of τ				
				Average estimate	Bias	Empirical SD	Average Robust SE	% CI coverage
B	100	11	(C,τ)	50.0	0.0	2.77	2.70	0.914
			(p_1,p_2,τ)	49.8	-0.2	3.67	2.21	0.862
		21	(C,τ)	49.9	-0.1	2.75	2.30	0.905
			(p_1,p_2,τ)	49.9	-0.1	3.02	1.84	0.837
	500	11	(C,τ)	50.0	0.0	1.12	1.28	0.928
			(p_1,p_2,τ)	49.9	-0.1	1.00	1.03	0.918
		21	(C,τ)	50.0	0.0	0.97	1.09	0.946
			(p_1,p_2,τ)	50.0	0.0	0.86	0.86	0.942

Table 3.11: Properties of estimates of τ from GEE estimation of (C,τ) and GEE estimation of (p_1,p_2,τ). The simplified variance formula $\hat{\text{Var}}_{\text{GF}}\left(S(t_j)\,|\,\Phi\right)$ was used in estimation of (C,τ) and multinomial variances (see text) of z_1 and z_2 were used in estimation of (p_1,p_2,τ).

M=number of markers. Markers are equally spaced from 0 to 100 cM.

Model B: penetrance vector = (0.001,0.001,0.2), p(D) = 0.01, $C = 0.483$, ($p_0 = 0.127$, $p_1 = 0.263$, $p_2 = 0.610$), $\tau = 50.0$.

% CI coverage = observed coverage of nominal 95% CIs for the disease gene location.

3.4 Discussion

In this chapter, several modifications of the GEE method for estimating the location of a single disease gene and the mean IBD sharing at this locus in ASPs were developed. It was shown that efficiency of the estimators is higher when a more accurate model for the covariance of the data is used. Furthermore, we demonstrated that by using the estimated probabilities of sharing 0, 1, or 2 alleles IBD at all markers as the observation for each ASP, IBD probabilities at the disease locus could be estimated by a GEE approach.

In section 3.1, alternative ways of specifying and calculating $\mathrm{Var}(S_i^*(t_m)|\Phi)$ were considered. The two approximations $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m)\,|\,\Phi\right)$ and $\hat{\mathrm{Var}}_{\mathrm{c}}\left(S(t_m)\,|\,\Phi\right)$ approximate $\mathrm{Var}(S_i^*(t_m)|\Phi)$ by $\mathrm{Var}(S_i(t_m)|\Phi)$. These two methods differ in the way they estimate one of the variance parameters, $p_0 = \Pr(S(\tau)=0|\Phi)$. $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m)\,|\,\Phi\right)$ is based on setting $p_0 = 0$, while $\hat{\mathrm{Var}}_{\mathrm{c}}\left(S(t_m)\,|\,\Phi\right)$ estimates p_0 using data at the nearest marker. In general, if a marker at location τ is not linked to the disease, $p_0 = ¼$, and if there is a disease gene at τ, $p_0 < ¼$. At a locus t_k linked to τ, $p_0 < p_{0k} < 1/4$. Therefore, whereas $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m)\,|\,\Phi\right)$ tends to underestimate $\mathrm{Var}(S_i(t_m)|\Phi)$, an approach based on estimating p_0 by IBD sharing at a nearby marker will tend to overestimate $\mathrm{Var}(S_i(t_m)|\Phi)$. However, using data at the nearest marker to estimate p_0 should provide improved estimates of $\mathrm{Var}(S_i(t_m)|\Phi)$ if a fairly dense marker map is used. It should also be noted that for non-fully informative data, $\mathrm{Var}(S_i^*(t_m)|\Phi)$ tends to be lower than $\mathrm{Var}(S_i(t_m)|\Phi)$. Thus, $\hat{\mathrm{Var}}_{\mathrm{c}}\left(S(t_m)\,|\,\Phi\right)$ will tend to overestimate the variance of non-fully informative data, but may work well in highly informative data when a fairly dense map is used. $\hat{\mathrm{Var}}_{\mathrm{GF}}\left(S(t_m)\mid\Phi\right)$, which tends to be lower than Var($S_i(t_m)|\Phi$), may, in practice, approximate the variance of non-fully informative data quite well.

Empirically estimating variances of S^* avoids problems associated with modeling the variance of potentially non-fully informative IBD sharing. Simulations demonstrated that empirically estimating variances of S^* leads to results comparable to those obtained by modeling the variance function. This finding has important implications for extensions of the GEE approach to localizing more than one disease gene in a region.

Although modeling the correlations of IBD sharing at linked markers can result in more efficient estimators, some computational costs are associated with more complex correlation structures. The AR-1 correlation structure that was incorporated into the one-locus GEE in section 3.2 was not difficult to implement because it has a simple inverse. Incorporating the correctly modeled correlation (equations 3.7) into the algorithm would require inversion of large matrices at each iteration, and may lead to slower convergence. However, more careful modeling of the correlation should be considered if the markers are not equally spaced, as the AR-1 correlation would not be as appropriate in that case.

In section 3.3 it was shown that by using the (estimated) IBD sharing probabilities at all markers as the observation for each ASP, the IBD sharing probabilities at the disease gene could be estimated in addition to the mean IBD sharing in ASPs at the disease gene. Although knowledge of p_j=Pr($S(\tau)$=$j|\Phi$) j = 0,1,2, still does not fully specify the underlying genetic model, these parameters may contain more information about the underlying genetic model than the single parameter C = E($S(\tau)|\Phi$)-1. Furthermore, for some underlying genetic models, by using more of the available data, this new parameterization can lead to improved estimation of the disease gene location in large samples.

The results presented in this chapter are not extensive, and are not intended for drawing comprehensive conclusions. Rather, in this chapter, several ideas were introduced, and small simulation studies were carried out to provide some indication about general properties of estimators resulting from certain modifications of the methods. The algorithms used for these simulations were not optimized. For instance, derivatives of the mean function with respect to the parameter τ are discontinuous, and no corrections for this were introduced into the algorithms for estimating $\eta = (p_1,p_2,\tau)$ in section 3.3. Further improvements of some of the methods discussed in this chapter could be achieved through more careful design of algorithms for solving the estimating equations. Nevertheless, these initial implementations did provide useful results allowing a general comparison of the effect of certain features of the one-locus GEE on the performance of the estimation procedure.

Chapter 4:

A Two-Locus Model for Estimation of Locations of Two Linked Disease Genes

Extending the work of Liang et al. (2001*a*), in this chapter we develop a model for the simultaneous localization of two susceptibility genes in one chromosomal region. We first derive an expression for mean allele sharing in affected sib pairs at each point across a chromosomal segment containing two susceptibility genes. With this expression for the mean allele sharing, generalized estimating equations (GEE) modeling can be used to localize both disease genes simultaneously. We develop an algorithm that uses identical-by-descent (IBD) sharing of marker alleles in affected sib pairs to estimate a set of parameters (δ_2) by solving the estimating equations:

$$\sum_{i=1}^{n}\left(\frac{\partial\mu_2(\delta_2)}{\partial\delta_2}\right)'\mathrm{Cov}^{-1}(S_i\mid\Phi)(S_i-\mu_2(\delta_2))=0\,, \tag{4.1}$$

where n is the number of families, S_i is the IBD sharing at all the markers for all the ASPs from the ith family, $\mu_2(\delta_2) = \mathrm{E}(S_i \mid \Phi)$, and Φ denotes ascertainment of an affected sib pair.

In subsequent sections, we derive the GEE method for estimating the locations of two linked disease genes. We first describe a model for IBD allele sharing at any position in a region containing two linked disease genes (section 4.1). Then, in section 4.2, we show how this model can be used in a GEE framework to estimate a set of parameters which includes the two gene locations. We conclude this chapter with a description of the implementation of this method (section 4.3).

4.1 A Model for Mean IBD Allele Sharing at a Locus Linked to Two Disease Genes

4.1.1 Expectation of IBD sharing at a locus *t* linked to two disease genes in terms of expected allele sharing at the two disease gene loci

Let τ_1 and τ_2 be the true locations of two linked trait loci ($\tau_1<\tau_2$), $S(t)$ be the number of alleles (0, 1, or 2) shared IBD by an ASP at locus t, and Φ denote the event that both sibs are affected. Define $C_1 = \mathrm{E}(S(\tau_1)|\Phi) - 1$ and $C_2 = \mathrm{E}(S(\tau_2)|\Phi) - 1$. (Calculation of these expectations for any two-locus model will be explained in section 5.2.2). Also let $\Psi_1 = \Psi_{t,\tau_1}$, $\Psi_2 = \Psi_{t,\tau_2}$, and $\Psi_3 = \Psi_{\tau_1,\tau_2}$ where Ψ is defined in the same way as in section 2.3, i.e. $\Psi = \theta^2 + (1-\theta)^2$, θ being the recombination fraction between two loci. Then under the assumptions of random mating, linkage equilibrium, generalized single ascertainment (Hodge and Vieland 1996), no interference, and equality of recombination fractions for males and females, $E(S(t)\mid\Phi)$ for any location *t* in a region with two linked disease genes can be derived as follows (further details are provided in appendix A).

First note that

$$E[S(t)\mid\Phi]=\Pr(S(t)=1\mid\Phi)+2\Pr(S(t)=2\mid\Phi)$$

Also,

$$\Pr(S(t)=j\mid\Phi)=\sum_{k=0}^{2}\sum_{l=0}^{2}\Pr(S(t)=j,S(\tau_1)=k,S(\tau_2)=l\mid\Phi)$$

$$=\sum_{k=0}^{2}\sum_{l=0}^{2}\Pr(S(t)=j\mid S(\tau_1)=k,S(\tau_2)=l,\Phi)\Pr(S(\tau_1)=k,S(\tau_2)=l\mid\Phi)$$

$$=\sum_{k=0}^{2}\sum_{l=0}^{2}\Pr(S(t)=j\mid S(\tau_1)=k,S(\tau_2)=l)\Pr(S(\tau_1)=k,S(\tau_2)=l\mid\Phi)$$

[assuming there are two linked genes at τ_1 and τ_2 and no other genes in the region, and assuming linkage equilibrium between markers and the trait loci so that genotypes at the trait loci and marker IBD sharing are independent given IBD sharing at the trait loci]

Case I: $t<\tau_1<\tau_2$

For $t<\tau_1<\tau_2$: $\Pr(S(t)=j\mid S(\tau_1)=k,S(\tau_2)=l)=\Pr(S(t)=j\mid S(\tau_1)=k)$ under the assumption of no interference, so that IBD sharing at consecutive loci behaves in a first-order Markov manner.

Therefore,

$$\Pr(S(t)=j\mid\Phi)=\sum_{k=0}^{2}\sum_{l=0}^{2}\Pr(S(t)=j\mid S(\tau_1)=k)\Pr(S(\tau_1)=k,S(\tau_2)=l\mid\Phi)$$

$$=\sum_{k=0}^{2}\Pr(S(t)=j\mid S(\tau_1)=k)\sum_{l=0}^{2}\Pr(S(\tau_1)=k,S(\tau_2)=l\mid\Phi)$$

$$= \sum_{k=0}^{2} \Pr\left(S(t) = j \mid S(\tau_1) = k\right) \Pr\left(S(\tau_1) = k \mid \Phi\right)$$

As the above formula shows, in order to calculate the probabilities Pr(S(t)=1|Φ) and Pr(S(t)=2|Φ), we need conditional probabilities of a sib pair sharing j alleles IBD at t given that they share k alleles IBD at τ_1. The matrix of conditional sharing probabilities in sibpairs Pr(S_A=i|S_B=j) i,j=0,1,2 for two linked loci A and B was given by Haseman and Elston (1972) and was reproduced in Table 3.7. Note that we have assumed equal recombination fractions in males and females so that $\theta_m = \theta_f = \theta$ and $\Psi = \theta^2 + (1-\theta)^2$. Using the conditional IBD sharing probabilities from Table 3.7, after simplification, it can be shown that:

$$E\left[S(t) \mid \Phi\right] = 1 + \left(2\Psi_1 - 1\right) C_1$$

where C_1=E[S(τ_1)|Φ]-1.

Case II: τ_1<τ_2<t

For τ_1<τ_2<t: $\Pr\left(S(t) = j \mid S(\tau_1) = k, S(\tau_2) = l\right) = \Pr\left(S(t) = j \mid S(\tau_2) = l\right)$.

Now similarly to case I, it can be shown that

$$E\left[S(t) \mid \Phi\right] = 1 + \left(2\Psi_2 - 1\right) C_2$$

where C_2=E[S(τ_2)|Φ]-1

Case III: τ_1<t<τ_2

For τ_1<t<τ_2:

$$\Pr\left(S(t) = j \mid S(\tau_1) = k, S(\tau_2) = l\right) = \frac{\Pr\left(S(t) = j, S(\tau_1) = k, S(\tau_2) = l\right)}{\Pr\left(S(\tau_1) = k, S(\tau_2) = l\right)}$$

$$= \frac{\Pr\left(S(\tau_2) = l \mid S(\tau_1) = k, S(t) = j\right)\Pr\left(S(\tau_1) = k, S(t) = j\right)}{\Pr\left(S(\tau_2) = l \mid S(\tau_1) = k\right)\Pr\left(S(\tau_1) = k\right)}$$

$$= \frac{\Pr\left(S(\tau_2) = l \mid S(t) = j\right)\Pr\left(S(t) = j \mid S(\tau_1) = k\right)\Pr\left(S(\tau_1) = k\right)}{\Pr\left(S(\tau_2) = l \mid S(\tau_1) = k\right)\Pr\left(S(\tau_1) = k\right)}$$

$$= \frac{\Pr\left(S(\tau_2) = l \mid S(t) = j\right)\Pr\left(S(t) = j \mid S(\tau_1) = k\right)}{\Pr\left(S(\tau_2) = l \mid S(\tau_1) = k\right)}$$

under the assumption of no interference.

Therefore,

$$E(S(t) \mid \Phi) = \Pr\left(S(t) = 1 \mid \Phi\right) + 2\Pr\left(S(t) = 2 \mid \Phi\right)$$

$$= \sum_{k=0}^{2}\sum_{l=0}^{2} \Pr\left(S(\tau_1) = k, S(\tau_2) = l \mid \Phi\right)\Pr\left(S(t) = 1 \mid S(\tau_1) = k, S(\tau_2) = l\right) +$$
$$2\sum_{k=0}^{2}\sum_{l=0}^{2} \Pr\left(S(\tau_1) = k, S(\tau_2) = l \mid \Phi\right)\Pr\left(S(t) = 2 \mid S(\tau_1) = k, S(\tau_2) = l\right)$$

$$= \sum_{k=0}^{2}\sum_{l=0}^{2} p_{kl} \frac{\Pr\left(S(\tau_2) = l \mid S(t) = 1\right)\Pr\left(S(t) = 1 \mid S(\tau_1) = k\right)}{\Pr\left(S(\tau_2) = l \mid S(\tau_1) = k\right)} +$$
$$2\sum_{k=0}^{2}\sum_{l=0}^{2} p_{kl} \frac{\Pr\left(S(\tau_2) = l \mid S(t) = 2\right)\Pr\left(S(t) = 2 \mid S(\tau_1) = k\right)}{\Pr\left(S(\tau_2) = l \mid S(\tau_1) = k\right)}$$

where $p_{kl} = \Pr\left(S(\tau_1) = k, S(\tau_2) = l \mid \Phi\right)$.

By substituting in expressions for the conditional sharing probabilities from Table 3.7, and using the properties:

$$\sum_{k=0}^{2}\sum_{l=0}^{2} p_{kl} = 1,$$

$$E(S(\tau_1) \mid \Phi) = \Pr\left(S(\tau_1) = 1 \mid \Phi\right) + 2\Pr\left(S(\tau_1) = 2 \mid \Phi\right)$$
$$= \sum_{l=0}^{2} \Pr\left(S(\tau_1) = 1, S(\tau_2) = l \mid \Phi\right) + 2\sum_{l=0}^{2} \Pr\left(S(\tau_1) = 2, S(\tau_2) = l \mid \Phi\right)$$
$$= \sum_{l=0}^{2} p_{1l} + 2\sum_{l=0}^{2} p_{2l}$$

and

$$\begin{aligned}
E(S(\tau_2)\mid\Phi) &= \Pr\left(S(\tau_2)=1\mid\Phi\right)+2\Pr\left(S(\tau_2)=2\mid\Phi\right)\\
&=\sum_{k=0}^{2}\Pr\left(S(\tau_1)=k,S(\tau_2)=1\mid\Phi\right)+2\sum_{k=0}^{2}\Pr\left(S(\tau_1)=k,S(\tau_2)=2\mid\Phi\right)\\
&=\sum_{k=0}^{2}p_{k1}+2\sum_{k=0}^{2}p_{k2}
\end{aligned}$$

it can be shown that:

$$\begin{aligned}
E(S(t)\mid\Phi) &= 1+\frac{\Psi_2\left(\Psi_2-1\right)\left(2\Psi_1-1\right)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}E(S(\tau_1)\mid\Phi)\\
&+\frac{\Psi_1\left(\Psi_1-1\right)\left(2\Psi_2-1\right)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}E(S(\tau_2)\mid\Phi)-\frac{\left(2\Psi_1-1\right)+\left(2\Psi_2-1\right)}{1+\left(2\Psi_1-1\right)\left(2\Psi_2-1\right)}
\end{aligned}$$

$$\begin{aligned}
&=1+\frac{\Psi_2\left(\Psi_2-1\right)\left(2\Psi_1-1\right)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}C_1\\
&+\frac{\Psi_1\left(\Psi_1-1\right)\left(2\Psi_2-1\right)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}C_2\\
&=1+\frac{\left(2\Psi_1-1\right)\left[1-\left(2\Psi_2-1\right)^2\right]}{1-\left(2\Psi_3-1\right)^2}C_1+\frac{\left(2\Psi_2-1\right)\left[1-\left(2\Psi_1-1\right)^2\right]}{1-\left(2\Psi_3-1\right)^2}C_2\,.
\end{aligned}$$

Thus, it has been shown that in a region containing two linked disease genes:

$$E(S(t)\mid\Phi)=$$

$$\begin{cases}
1+\left(2\Psi_1-1\right)\cdot C_1 & \text{for } t<\tau_1<\tau_2\\
1+\dfrac{\left(2\Psi_1-1\right)\left[1-\left(2\Psi_2-1\right)^2\right]}{1-\left(2\Psi_3-1\right)^2}C_1+\dfrac{\left(2\Psi_2-1\right)\left[1-\left(2\Psi_1-1\right)^2\right]}{1-\left(2\Psi_3-1\right)^2}C_2 & \text{for } \tau_1<t<\tau_2\\
1+\left(2\Psi_2-1\right)\cdot C_2 & \text{for } \tau_1<\tau_2<t
\end{cases}\tag{4.2}$$

As in the single-locus model derived by Liang et al. (2001*a*), the above formula for E(S(t)|Φ) is robust in the sense that it holds regardless of the mode of inheritance. The parameters C_1 and C_2 depend on the underlying genetic mechanism, and E(S(t)|Φ) depends on the genetic model only through the parameters C_1 and C_2 and the position of t relative to τ_1 and τ_2. No assumptions are made about the existence of other unlinked trait genes.

Note that when t is not linked to either of the two disease genes, then $\Psi_1 = \Psi_{t,\tau_1} = 0.5$, $\Psi_2 = \Psi_{t,\tau_2} = 0.5$, and E($S$($t$)|Φ) = 1 as expected under the null hypothesis of no linkage to any disease genes. Also, if τ_1 and τ_2 are not linked, and t is linked only to τ_1, then $\Psi_2 = \Psi_{t,\tau_2} = 0.5$, $\Psi_3 = \Psi_{\tau_1,\tau_2} = 0.5$, and $E(S(t) \mid \Phi) = 1 + (2\Psi_1 - 1) \cdot C_1$ for all t. That is, the two-locus model reduces to the one-locus model of Liang et al. (2001*a*) in this special case.

Figure 4.1 shows the E(S(t)|Φ) curve for a chromosomal segment containing two disease genes under two different genetic models specified by the matrix of joint penetrances and allelic frequencies. One is a dominant by dominant epistatic model and the other is a codominant by dominant additive model. Haldane's mapping function was used to translate recombination fraction to genetic map distance [$\theta_{t_1,t_2} = \left(1 - e^{-0.02|t_1 - t_2|}\right)/2$], and thus re-write $\mu_2(t)$ = E(S(t)|Φ) in terms of t and δ_2=(C_1,C_2,τ_1,τ_2) rather than in terms of C_1, C_2, and Ψ's or θ's. We use $\mu_2(t;\delta_2)$ to denote this new form of the two-locus model mean function for all t.

$$\mu_2(t;\delta_2) =$$

$$\begin{cases} 1+e^{-.04|t-\tau_1|}\cdot C_1 & \text{for } t<\tau_1<\tau_2 \\ 1+C_1e^{-.04(t-\tau_1)}\dfrac{1-e^{-.08(\tau_2-t)}}{1-e^{-.08(\tau_2-\tau_1)}}+C_2e^{-.04(\tau_2-t)}\dfrac{1-e^{-.08(t-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}} & \text{for } \tau_1<t<\tau_2 \\ 1+e^{-.04|t-\tau_2|}\cdot C_2 & \text{for } \tau_1<\tau_2<t \end{cases} \qquad (4.3)$$

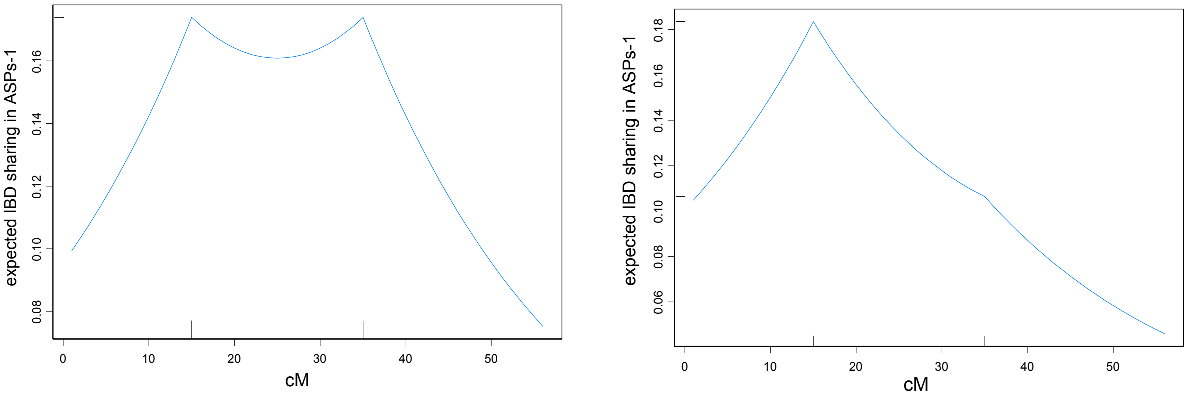


Figure 4.1: Plots of E[$S(t)|\Phi$] for two different genetic models with two linked disease genes.

A) A dominant*dominant epistatic model given by:

Penetrances:		Locus 1		
		aa	aA	AA
Locus 2	bb	.05	.475	.475
	bB	.475	.90	.90
	BB	.475	.90	.90

Allele Frequencies: Pr(A) = Pr(B) = 0.015.

Positions of two disease loci: τ_1 = 15cM, τ_2 = 35cM

B) A codominant*dominant additive model given by

Penetrances:		Locus 1		
		aa	aA	AA
Locus 2	bb	.05	.475	.475
	bB	.225	.65	.65
	BB	.475	.90	.90

Allele Frequencies: Pr(A) = Pr(B) = 0.015.

Positions of two disease loci: τ_1 = 15cM, τ_2 = 35cM

4.1.2 Expectation of IBD sharing at a locus linked to two disease genes in terms of an alternative set of parameters

Consider the re-parameterization:

$$C_1 = C_1^* + C_2^* e^{-.04(\tau_2-\tau_1)} \text{ and } C_2 = C_2^* + C_1^* e^{-.04(\tau_2-\tau_1)}.$$

That is:

$$C_1^* = \frac{C_1 - C_2 e^{-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}} \text{ and } C_2^* = \frac{C_2 - C_1 e^{-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}}.$$

Then for $t<\tau_1<\tau_2$:

$$\mu_2^*(t;\delta_2^*) = 1 + e^{-.04(\tau_1-t)}\left[C_1^* + C_2^* e^{-.04(\tau_2-\tau_1)}\right] = 1 + C_1^* e^{-.04(\tau_1-t)} + C_2^* e^{-.04(\tau_2-t)},$$

for $\tau_1<\tau_2<t$:

$$\mu_2^*(t;\delta_2^*) = 1 + e^{-.04(t-\tau_2)}\left[C_2^* + C_1^* e^{-.04(\tau_2-\tau_1)}\right] = 1 + C_2^* e^{-.04(t-\tau_2)} + C_1^* e^{-.04(t-\tau_1)},$$

and for $\tau_1<t<\tau_2$:

$$\begin{aligned}\mu_2^*(t;\delta_2^*) = 1 + &\frac{e^{-.04(t-\tau_1)} - e^{-.04(\tau_2-t)-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}}\left[C_1^* + C_2^* e^{-.04(\tau_2-\tau_1)}\right] + \\ &\frac{e^{-.04(\tau_2-t)} - e^{-.04(t-\tau_1)-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}}\left[C_2^* + C_1^* e^{-.04(\tau_2-\tau_1)}\right]\end{aligned}$$

$$\begin{aligned}&= 1 + C_1^*\left[\frac{e^{-.04(t-\tau_1)} - e^{-.04(\tau_2-t)-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}} + e^{-.04(\tau_2-\tau_1)}\frac{e^{-.04(\tau_2-t)} - e^{-.04(t-\tau_1)-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}}\right] + \\ &C_2^*\left[e^{-.04(\tau_2-\tau_1)}\frac{e^{-.04(t-\tau_1)} - e^{-.04(\tau_2-t)-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}} + \frac{e^{-.04(\tau_2-t)} - e^{-.04(t-\tau_1)-.04(\tau_2-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}}\right]\end{aligned}$$

$$= 1 + C_1^* e^{-.04(t-\tau_1)} + C_2^* e^{-.04(\tau_2-t)}$$

Therefore,

$$\mu_2^*(t) = E(S(t)\,|\,\Phi) = 1 + C_1^* e^{-.04|t-\tau_1|} + C_2^* e^{-.04|t-\tau_2|} \quad \text{for all } t.$$

Note that if $C_1^* = 0$ then $\mu_2(t) = \mu_1(t;C_2^*,\tau_2) = E(S(t)\,|\,\Phi) = 1 + C_2^* e^{-.04|t-\tau_2|}$.

Similarly, if $C_2^* = 0$ then $\mu_2(t) = \mu_1(t; C_1^*, \tau_1) = E(S(t) \mid \Phi) = 1 + C_1^* e^{-.04|t-\tau_1|}$.

That is, the two-locus model parameterized by $\delta_2^*=(C_1^*, C_2^*, \tau_1, \tau_2)$ reduces to the one-locus model of Liang et al. (2001*a*) if either of the C^* parameters equals 0. This implies that each C^* parameter represents the excess sharing at a location due to the gene at that location, and can thus be seen as the "effect size" of that particular gene; whereas the *C* parameters (C_1 and C_2) represent the mean excess IBD sharing at the location of one of the disease genes, which is increased by effects due to both disease genes.

Although the interpretation of C_1 and C_2 is clear (i.e. one less than the expected IBD sharing at the locations of the two disease genes, τ_1 and τ_2, respectively), interpretation of C_1^*

Model					D (cM)	C_1	C_2	C_1^*	C_2^*
Penetrance Matrix				Pr(A), Pr(B)					
	aa	aA	AA	0.2, 0.2	10000	0.083	0.083	0.083	0.083
bb	0.0	0.1	0.2		100	0.085	0.085	0.083	0.083
bB	0.1	0.2	0.3		50	0.095	0.095	0.083	0.083
BB	0.2	0.3	0.4		30	0.108	0.108	0.083	0.083
					10	0.139	0.139	0.083	0.083
					1	0.163	0.163	0.083	0.083
	aa	aA	AA	0.1,0.1	10000	0.336	0.336	0.336	0.336
bb	.001	.001	.001		100	0.341	0.341	0.335	0.335
bB	.001	.200	.200		50	0.371	0.371	0.327	0.327
BB	.001	.200	.200		30	0.411	0.411	0.316	0.316
					10	0.491	0.491	0.294	0.294
					1	0.548	0.548	0.279	0.279

Table 4.1: C_1^* and C_2^* for two genetic models.
D = distance between two genes, in centiMorgans

and C_2^* is not as obvious. Empirical results suggest that when the two-locus disease model is additive on the penetrance scale (see examples in Table 4.1), C_1^* and C_2^* equal the excess IBD sharing at the two disease genes under a model with the same penetrance matrix and allele frequencies, but with the two disease genes unlinked. In this case, C_1^* and C_2^* can be interpreted as effects on IBD sharing due to gene 1 and gene 2, respectively. For other non-additive models, interpretation of C_1^* and C_2^* is not as straightforward.

4.1.3 Variance of IBD sharing at a locus linked to two disease genes

Details of the derivations of variance functions presented in this section are shown in the appendix. The variances of IBD sharing values at the two disease loci are:

$$\mathrm{Var}\left(S(\tau_1)\,|\,\Phi\right) = C_1 - C_1^2 + 2\Pr\left(S(\tau_1)=0\,|\,\Phi\right) \text{ and}$$

$$\mathrm{Var}\left(S(\tau_2)\,|\,\Phi\right) = C_2 - C_2^2 + 2\Pr\left(S(\tau_2)=0\,|\,\Phi\right).$$

At fully informative markers, the variance at a marker t depends on the location of t in relation to the two disease loci.

For $t < \tau_1 < \tau_2$: $\mathrm{Var}\left(S(t)\,|\,\Phi\right) = \left(2\Psi_{t,\tau_1} - 1\right)\left[\mathrm{Var}\left(S(\tau_1)\,|\,\Phi\right) - 0.5\right] + 0.5$.

For $\tau_1 < \tau_2 < t$: $\mathrm{Var}\left(S(t)\,|\,\Phi\right) = \left(2\Psi_{t,\tau_2} - 1\right)\left[\mathrm{Var}\left(S(\tau_2)\,|\,\Phi\right) - 0.5\right] + 0.5$.

For $\tau_1 < t < \tau_2$: $\mathrm{Var}\left(S(t)\,|\,\Phi\right) = E\left(S(t)\,|\,\Phi\right) - \left[E\left(S(t)\,|\,\Phi\right)\right]^2 + 2\Pr\left(S(t)=2\,|\,\Phi\right)$.

$E(S(t)|\Phi)$ is given in equation 4.3, and is a function of C_1, C_2, τ_1, and τ_2. $\Pr(S(t)=2|\Phi)$ is a function of the distances between t, τ_1 and τ_2, as well as the nine joint sharing probabilities p_{ij} $= \Pr(S(\tau_1)=i, S(\tau_2)=j|\Phi)$ $i,j=0,1,2$ ($\sum_{i=0}^{2}\sum_{j=0}^{2} p_{ij} = 1$) which depend on the disease model through penetrances and disease allele frequencies at the disease loci, and on the distance between τ_1 and τ_2. It can be shown that:

$$\begin{aligned}\Pr\left(S(t)=2\,|\,\Phi\right) &= \frac{\left(1-\Psi_1\right)^2\left(1-\Psi_2\right)^2 p_{00}}{\Psi_3{}^2}+\frac{\left(1-\Psi_1\right)^2 2\Psi_2\left(1-\Psi_2\right)p_{01}}{2\Psi_3\left(1-\Psi_3\right)}+\frac{\left(1-\Psi_1\right)^2\Psi_2^2 p_{02}}{\left(1-\Psi_3\right)^2}\\ &+\frac{\Psi_1\left(1-\Psi_1\right)\left(1-\Psi_2\right)^2 p_{10}}{\Psi_3\left(1-\Psi_3\right)}+\frac{\Psi_1\left(1-\Psi_1\right)2\Psi_2\left(1-\Psi_2\right)p_{11}}{1-2\Psi_3\left(1-\Psi_3\right)}+\frac{\Psi_1\left(1-\Psi_1\right)\Psi_2^2 p_{12}}{\Psi_3\left(1-\Psi_3\right)}\\ &+\frac{\Psi_1^2\left(1-\Psi_2\right)^2 p_{20}}{\left(1-\Psi_3\right)^2}+\frac{\Psi_1^2 2\Psi_2\left(1-\Psi_2\right)p_{21}}{2\Psi_3\left(1-\Psi_3\right)}+\frac{\Psi_1^2\Psi_2^2 p_{22}}{\Psi_3{}^2}\end{aligned}$$

4.2 Estimation of the Locations of Two Trait Loci

4.2.1 Parameter estimation

In the two-locus model, $\mu_2(t;\delta_2)$ is characterized parametrically by four parameters: C_1, C_2, τ_1, and τ_2. If a sample of n independent ASPs were collected, and IBD sharing for all ASPs at all marker loci were known, one could obtain estimates of the parameters $\delta_2 = (C_1, C_2, \tau_1, \tau_2)$ by solving the GEE:

$$\sum_{i=1}^{n}\left(\frac{\partial\mu_2(\delta_2)}{\partial\delta_2}\right)'\mathrm{Cov}^{-1}(S_i\,|\,\Phi)(S_i-\mu_2(\delta_2))=0$$

where $S_i = (S_i(t_1),\ldots,S_i(t_M))' = (S_{i1},\ldots,S_{iM})'$ and $\mu_2(\delta_2) = (\mu_2(t_1;\delta_2),\ldots, \mu_2(t_M;\delta_2))'$. If there are two disease genes in the region, solving these equations for δ_2 should provide consistent estimates of the parameters and the variances of the estimates since it has been shown that under the two-locus model E$[S_i]$ $= \mu_2(\delta_2)$ for ASPs.

However, in most cases, exact IBD sharing is not known for all ASPs at all markers. Rather, estimated IBD sharing can be obtained, for example, using the multipoint methods incorporated in the software program GENEHUNTER (Kruglyak et al. 1996). These imputed IBD statistics, $S_i^*(t_m)$, use the multipoint marker information, allele frequencies, and the assumed recombination fractions between markers, and estimate $S_i(t_m)$ by considering all possible IBD configurations consistent with the observed marker information:

$$S_i^*(t_m) = \sum_{l=0}^{2} l \cdot \Pr(S_i(t_m) = l \mid Y_i; \mathrm{H}_0)\,.$$

Following the methods of Liang et al. (2001*a*) for the single-locus-model, estimates of the two-locus-model parameters $\delta_2 = (C_1, C_2, \tau_1, \tau_2)$ can be obtained by substituting the estimated IBD allele sharing $S_i^* = (S_i^*(t_1),\ldots, S_i^*(t_M))' = (S_{i1}^*, S_{i2}^*,\ldots, S_{iM}^*)'$ for S_i in the GEE shown above and solving the resulting equations:

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_2(\delta_2)}{\partial \delta_2}\right)' \mathbf{V}_i^{-1}(S_i^* - \mu_2(\delta_2)) = 0 \qquad (4.4)$$

where $\mathbf{V}_i$ is the working covariance matrix for the data, S_i^*. In the next two subsections, the mean and covariance functions required to solve the GEE, $\mathrm{E}(S_i^*|\Phi)$ and $\mathrm{Cov}(S_i^*|\Phi)$, are described.

4.2.2 Specification of E[S_i^*|Φ]

$\mathrm{E}[S_i|\Phi] = \mu_2(\delta_2)$ was derived in section 4.1 (equation 4.3). Parameter estimates and their estimated precision will be valid as long as $\mathrm{E}[S_i^*|\Phi] = \mu_2(\delta_2)$. Assuming fully informative markers, we have $\mathrm{E}[S_i^*|\Phi] = \mathrm{E}[S_i|\Phi] = \mu_2(\delta_2)$ since at any fully informative marker m, $S_{im}^* = S_{im}$. At non-fully informative markers linked to one or more disease genes, S_{im}^* will tend to underestimate S_{im}, because $S_{im}^* = \mathrm{E}[S_{im}|Y_i,\mathrm{H}_0]$ is determined under the null hypothesis of no linkage to any disease genes. That is, S_{im}^* is the expected sharing at marker

m, conditional on all the marker data, computed without taking the affected status of the sib pairs into account. However, when the markers are highly (if not fully) informative, the expression will still hold approximately: $\mathrm{E}[S_i^*|\Phi] \approx \mu_2(\delta_2)$.

4.2.3 Specification of Cov[$S_i^*|\Phi$]

When markers are fully informative for the ith family, the complete-data vector S_i can be used in the GEE (equation 4.1). The diagonal elements of the complete-data covariance matrix $\mathrm{Cov}(S_i|\Phi)$, that is $\mathrm{Var}(S_i(t_m)|\Phi)$, were derived in section 4.1.3. When marker IBD sharing is not known with certainty because markers are not fully informative, marker IBD sharing can be determined probabilistically. The estimated IBD sharing, S_i^*, is then used in place of the true IBD sharing, S_i. As was discussed in chapter 3 for the one-locus model, if $\mathrm{Cov}(S_i^*|\Phi)$ is unknown, we may substitute it with the working covariance matrix, $\mathbf{V}_i = \mathrm{A}_i^{1/2}\mathrm{R}_{0i}(\alpha)\mathrm{A}_i^{1/2}$ where A_i is a $M \times M$ diagonal matrix of variances of $S_{i\mathrm{m}}^*$, and $\mathrm{R}_{0i}(\alpha)$ is the $M \times M$ working correlation matrix for S_i^* fully specified by a vector of unknown parameters, α.

In our implementation of the two-locus model GEE, we assume an independence correlation structure by setting $\mathrm{R}_{0i}(\alpha)$ equal to the identity matrix. That is, we assume that the correlation of IBD sharing at any two markers for the same sib pair is 0, and correlation of IBD sharing across sibpairs from the same family is also 0, and rely on the empirical component of the robust sandwich estimator to capture the correlation information. Therefore, $\mathbf{V}_i$ is a $M\times M$ diagonal matrix with $\mathrm{Var}(S_i^*(t_m)|\Phi)$, $m=1,\ldots,M$ on the diagonal.

$\mathrm{Var}(S_i^*(t_m)|\Phi)$ could be approximated by $\mathrm{Var}(S_i(t_m)|\Phi)$. In the one-locus model $\mathrm{Var}(S_i(t_m)|\Phi)$ is parametrized by C, τ, and one additional parameter, $\Pr(S(\tau)=0|\Phi)$. In the original version of Genefinder, Liang et al. (2001*a*) implemented an approximation such that

Var($S_i(t_m)$|Φ) was parameterized by C and τ only. Alternatively, as discussed in chapter 3 (section 3.1), Var($S_i^*(t_m)$|Φ) can be estimated empirically, for example by:

$$\hat{\mathrm{Var}}(S_i^*(t_m) \mid \Phi) = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_i^*(t_m) - \bar{S}^*(t_m)\right)^2 \text{, where } \bar{S}^*(t_m) = \frac{\sum_{i=1}^{n} S_i^*(t_m)}{n} \text{ ,}$$

n = number of ASPs.

Our simulations (section 3.1.4) have shown that for a range of simulation settings considered, empirically estimating variances of S^* at the markers led to results comparable to those obtained by modeling the variance using the approximate variance function implemented in Genefinder.

Based on the results for the one-locus model, we concluded that using empirical variances is likely to provide good estimates, and thus, in our implementation of the two-locus-model GEE, we also empirically estimate Var($S_i^*(t_m)$|Φ) m=1,…,M. For the two-locus model there is a stronger motivation for using empirical rather than model-based estimates of Var($S_i^*(t_m)$|Φ). As is shown in section 4.1.3, Var($S_i(t_m)$|Φ) is a function of the unknown joint sharing probabilities Pr($S(\tau_1)=k,S(\tau_2)=l$|Φ) (k=0,1,2; l=0,1,2) and the genetic distance between t and the two disease loci. Thus, Var($S_i(t_m)$|Φ) is a function of up to eight new parameters that do not contribute to the mean function and are therefore not estimated by the GEE (equation 4.1). These parameters could themselves be estimated by a GEE procedure (e.g. Prentice 1988). However, because of the large number of additional parameters that would need to be estimated, combined with the evidence about the adequacy of the empirical variances from the one-locus model, we decided not to model the variances parametrically, but rather estimate them empirically. We suspect that this approach may become problematic if a very large number of markers is used in the analysis. In that case, more Var($S_i^*(t_m)$|Φ) values would need to be estimated, which could lead to instability of estimates and their standard errors.

Under mild regularity conditions, a solution to the generalized estimating equations provides consistent estimates of the parameters δ_2 even when the working covariance matrix for the data, $\mathbf{V}_i$, is incorrectly specified (Liang and Zeger 1986). Furthermore, the robust sandwich estimator (Huber 1967; White 1982; Liang and Zeger 1986) gives consistent estimates of the variances of parameter estimates even when the correlation between observations is not modeled correctly. In this implementation, we assume an independence correlation structure and empirically estimate the diagonal elements of the covariance matrix, $\mathrm{Var}(S_i^*(t_m)|\Phi)$.

4.2.4 Estimation of the covariance of parameter estimates and construction of confidence intervals

By large-sample quasi-likelihood theory, solving the generalized estimating equations results in consistent and asymptotically normally distributed parameter estimates (Liang and Zeger 1986). Furthermore, although in small samples downward bias of standard errors obtained from the robust variance estimator has been observed in a number of situations, (e.g. Pan 2001), in large samples variances (and covariances) of the estimates are consistently estimated by the robust (sandwich) variance-covariance matrix (Liang and Zeger 1986):

$$\mathbf{C\hat{o}v}_R(\hat{\delta}_2) = \left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \left[\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} (S_i^* - \mu_2(\hat{\delta}_2))(S_i^* - \mu_2(\hat{\delta}_2))^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right] \left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1}$$

where $\hat{\mathbf{D}}_i = \dfrac{\partial \mu_2(\hat{\delta}_2)}{\partial \delta_2}$ and $\hat{\mathbf{V}}_i$ is the estimated working covariance matrix described in section 4.2.3.

Thus, asymptotic 100(1-2α)% confidence intervals (CIs) for the location estimates can be constructed as:

estimate ± $z_{\alpha}\cdot se$

where z_{α} is the 100(1-α)th percentile of the standard normal distribution, and *se* is the robust estimate of the standard error of the parameter estimate.

As with the single-locus model, estimates of the disease gene locations, τ_1 and τ_2, are of primary interest. Values of C_1 and C_2 depend on the underlying genetic model (penetrance matrix and allele frequencies at the two disease genes) and the distance between the disease genes. Although estimates of C_1 and C_2 are of limited use in the sense that knowledge of their values does not identify the underlying genetic model, their magnitude influences the accuracy and precision with which we can estimate the disease gene locations. As Liang et al. (2001*a*) noted for the one-locus model, information for localizing disease genes is thus reduced in the presence of oligogenic inheritance and/or genetic heterogeneity.

4.3 Implementation and Performance of the Algorithm (Numerical Issues/Convergence Problems)

We implemented a solution to the two-locus model GEE in a FORTRAN program with a modified fisher's scoring algorithm similar to the one used in GENEFINDER (Liang et al. 2001*a*). We obtain estimates of the parameters in our mean model by solving the GEE (equation 4.4). Parameter estimates are obtained by an iterative procedure:

$$\delta_{2,j} = \delta_{2,j-1} + \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_2(\delta_2)}{\partial \delta_2} \right)' \mathbf{V}_i^{-1} \left(\frac{\partial \mu_2(\delta_2)}{\partial \delta_2} \right) \right]_{\hat{\delta}_{2,j-1}}^{-1} \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_2(\delta_2)}{\partial \delta_2} \right)' \mathbf{V}_i^{-1} (S_i^* - \mu_2(\delta_2)) \right]_{\hat{\delta}_{2,j-1}}$$

beginning with some starting values $\delta_{2,0}$. For a scalar parameter δ, we say that the algorithm has converged to a solution providing an estimate when δ_j=$\delta_{j\text{-}1}$+/-ε for some small ε. For vector valued parameters of dimension p>1, there are many ways to define convergence, all of which conclude convergence when the previous estimate of δ is sufficiently close to the current estimate in some sense. We consider the algorithm to have converged when each of the p parameters changes during the last iteration by a value smaller than a preset threshold (0.001 for C_1 and C_2, and 0.005 for τ_1 and τ_2)

The algorithms for both the one-locus model and the two-locus model fail to converge for some data sets. This is not an unusual problem. For instance, in an application of GEEs to clustered polytomous data, O'Hara Hines (1998) reported that the algorithm sometimes failed to converge, particularly when more complex within-cluster correlation structures were assumed for the data. Lipsitz et al. (1994) also discussed such convergence problems, especially for small sample sizes. In our application of the GEE, the convergence rates were quite dependent on the underlying genetic model. Although we assumed the very simple independence correlation structure, the algorithm had difficulties converging in a substantial proportion of randomly generated data sets under certain genetic models, while for other models convergent solutions could be obtained in nearly 100% of generated data sets. Generally, convergence rates improved as the effect size of the genes (C parameters), distance between the two genes, or the number of affected sib pairs increased.

Different types of non-convergence occurred, including situations where the parameter estimates "bounced" around in a relatively small range of values, one or more of the parameters went out of bounds, or the parameters oscillated between two or more sets of values in a periodic manner. This type of "periodic convergence" to two points near the real solution appeared to be most frequent. In addition, in some data sets a solution was achieved through convergence of the algorithm, however, variance estimates for some parameters were not computed because the information matrix at the solution was (nearly) singular.

In order to assess properties of estimators such as bias, it is important to obtain a solution for a large proportion of generated data sets. Therefore, we implemented two strategies to obtain approximate solutions in data sets for which convergence was not initially attained. First of all, if a solution was not attained within a prescribed number of iterations, the "step size" was reduced, i.e. the procedure continues with:

$$\delta_{2,j} = \delta_{2,j-1} + \frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{\partial\mu_2(\delta_2)}{\partial\delta_2}\right)' \mathbf{V}_i^{-1}\left(\frac{\partial\mu_2(\delta_2)}{\partial\delta_2}\right)\right]^{-1}_{\hat{\delta}_{2,j-1}} \left[\sum_{i=1}^{n}\left(\frac{\partial\mu_2(\delta_2)}{\partial\delta_2}\right)' \mathbf{V}_i^{-1}(S_i^* - \mu_2(\delta_2))\right]\Bigg|_{\hat{\delta}_{2,j-1}}$$

The step size is further reduced a number of times if necessary. If the algorithm still fails to converge following the step-size reduction, a grid search is performed in the vicinity of the last few δ_j's if they do not span a very large area. When the grid-search is performed, the solution is chosen to be the set of parameters (among those considered) which gives the smallest sum of absolute values of the p components of the quasi-score vector. We found that these two simple procedures can yield estimates in a large proportion of "non-convergent" data sets. Furthermore, properties of these estimates seem comparable to those obtained by convergence without resorting to step-size reduction or grid-search, when it is possible.

Chapter 5

Assessment of the Estimation Method by Simulation

The new method to localize two linked trait loci described in chapter 4 was evaluated by simulation. The estimation method was first applied to two simulated data sets in which two linked loci influenced the probability of disease. These simulated examples, described in section 5.1, provide a more thorough view of individual results, such as visual comparison of NPL plots and plots of curves given by the GEE solutions. We then performed a more extensive simulation study to evaluate properties of the estimates, including bias of parameter estimators and their standard errors, efficiency, and confidence interval coverage for a range of genetic models and linkage study designs. In section 5.2 we describe the simulation study methods, in particular the generation of fully informative IBD sharing data for a sample of affected sib-pairs. In section 5.3 we report the results of several sets of simulations designed to investigate factors that influence the behaviour of the two-locus GEE estimators.

5.1 Illustration of Localization of Two Linked Disease Genes through Simulated Examples

5.1.1 Simulated example 1: Data Generated with GASP

For the first simulated example, quantitative trait data were generated using GASP (Wilson et al. 1996) and dichotomized using a threshold to produce disease status. The quantitative trait was influenced jointly by two linked loci. Both disease loci were assumed to be biallelic, with allele frequencies of 0.7 and 0.3. A liability threshold model was used, with each disease locus contributing 35% of the quantitative trait variability. The quantitative trait had a standard normal distribution, and a threshold of 1.28 was used to define individuals as affected. Genotypes at 15 linked markers were generated, covering about 30cM, assuming two trait loci at approximately 9.9 and 21.8 cM. Nuclear families with at least two affected sibs were selected for analysis. In total, there were 37 nuclear families in the analyzed sample, with 41 ASPs. GENEHUNTER 2.0 (Kruglyak et al. 1996) was used to compute estimates of IBD sharing for all ASPs and the NPL statistic at all markers (see Figure 5.1a). The GEE approaches described were then used to estimate positions of the disease loci and the expected allele sharing in ASPs at the disease loci assuming one-locus and two-locus models.

By fitting a one-locus model (Figure 5.1b, pg 91, solid line), τ was estimated to be 15.97 with a standard error of 4.59 (95% CI for τ: 7.0-25.0). C was estimated at 0.28 with a standard error of 0.15. By fitting a two-locus model (Figure 5.1b, pg 91, dashed line), τ_1 and τ_2 were estimated to be 8.20 (s.e. 3.28; 95% CI 1.8-14.6) and 24.35 (s.e. 5.57; 95% CI 13.4-

35.3) respectively. The estimates of C_1 and C_2 were 0.22 and 0.23 with standard errors of 0.13 and 0.14 respectively.

5.1.2 Simulated example 2: Data Generated with Allegro

For the second simulated example, the software Allegro (Gudbjartsson et al. 2000) was used to generate the data, under a locus heterogeneity model with two linked disease loci. Genotype data were generated for one set of nuclear families in which the disease locus was at position τ_1, and for another set of nuclear families in which the disease locus was at a different position, τ_2. Genotype data for both sets of families were generated at the same marker loci (i.e. on the same map) conditional on the two siblings being affected. The two sets of families were then combined and analyzed as one sample. A map of 19 equally spaced markers with a recombination fraction of 0.03 between consecutive markers, covering approximately 56 cM in total, was used. The sample consisted of 80 nuclear families with two affected sibs from each of two populations, for a total of 160 ASPs. In the first population, there was a diallelic disease susceptibility gene at 15 cM, with a disease allele frequency of 0.015 and penetrance vector (0.05,0.90,0.90). In the second population there was a diallelic disease susceptibility gene at 35cM, with the same allele frequencies and penetrances as the disease susceptibility gene in the first population. Thus, there were two disease genes, at positions $\tau_1 = 15.0$ and $\tau_2 = 35.0$, with equal contributions to disease susceptibility. The data was analyzed as in the other examples, that is with GENEHUNTER 2.0 (Kruglyak et al. 1996) followed by the two GEE approaches.

The NPL plot and average IBD sharing for this example are shown in Figure 5.2 (pg 92). The plot shows a wide NPL peak. Although the NPL score is high, reaching a maximum of 6.10 at 27.84cM, the NPL exceeds 3 for a very large chromosomal region, making it difficult to estimate the position(s) of potential disease gene(s). Furthermore, the wide peak suggests the possibility of more than one susceptibility gene in the region. Fitting a one-locus model leads to an estimated τ of 23.72 (s.e. 1.70; 95% CI: 20.4-27.1), and an estimated C of 0.432 (s.e. 0.059). By fitting a two-locus model the positions of the two disease genes are estimated at 13.97 (s.e. 2.03; 95% CI 10.0-17.9) and 35.07 (s.e. 2.65; 95% CI 29.9–40.3). The estimates of C_1 and C_2 are 0.346 and 0.329 respectively, with standard errors 0.059 and 0.052.

5.1.3 Summary of simulated examples

In both simulated examples the two-locus GEE method provides fairly accurate point estimates of the two disease gene locations. The method seems particularly valuable in the second example, when the disease genes are further apart. In this example the location confidence interval from the one-locus model does not include either of the two true disease gene locations, while estimates from the two-locus model provide two non-overlapping confidence intervals that include the two disease gene locations.

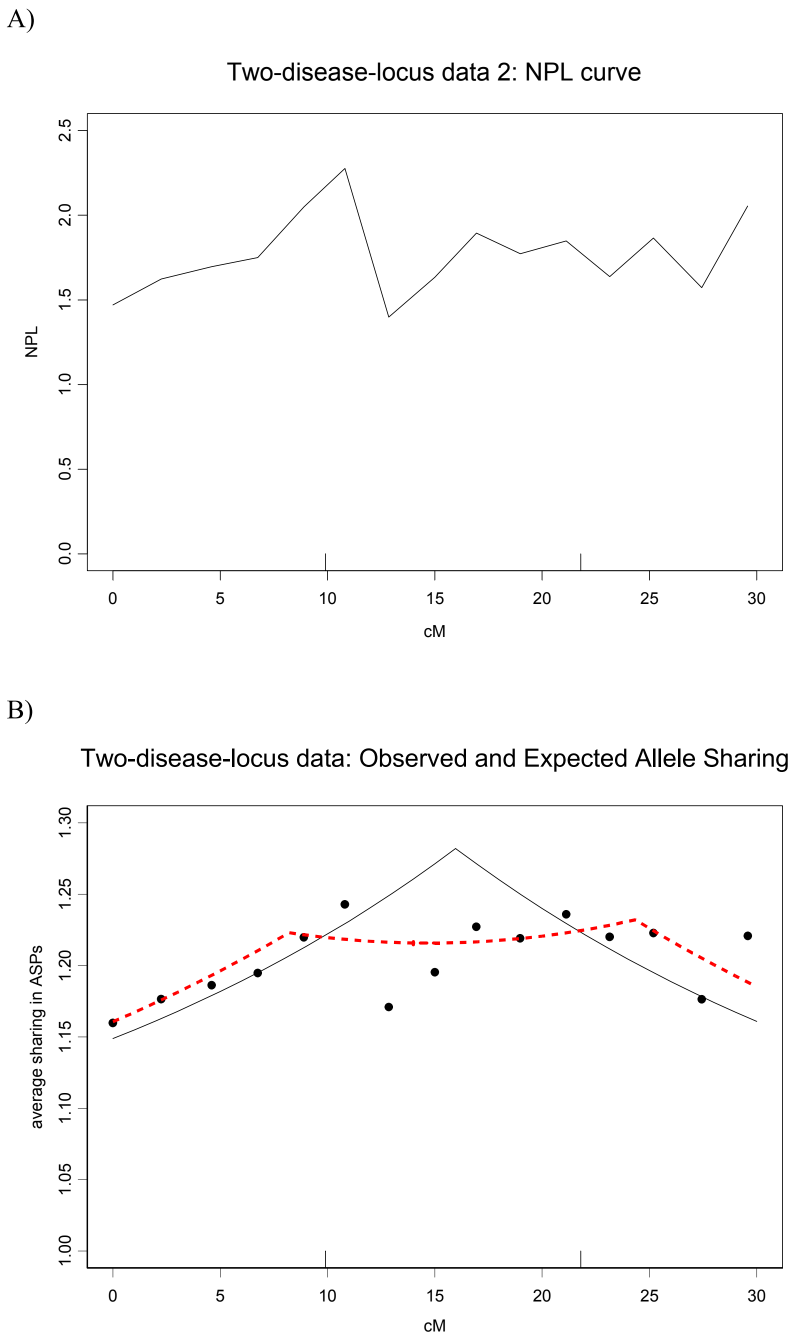


Figure 5.1: Plots for simulated example 1. (a) NPL plot; (b) observed average marker IBD allele-sharing in the sample of ASPs (points) with one-locus model fit (solid line) and two-locus model fit (dashed line). True disease gene locations are shown along the x-axis. The generating model under two linked disease loci is described in the text.

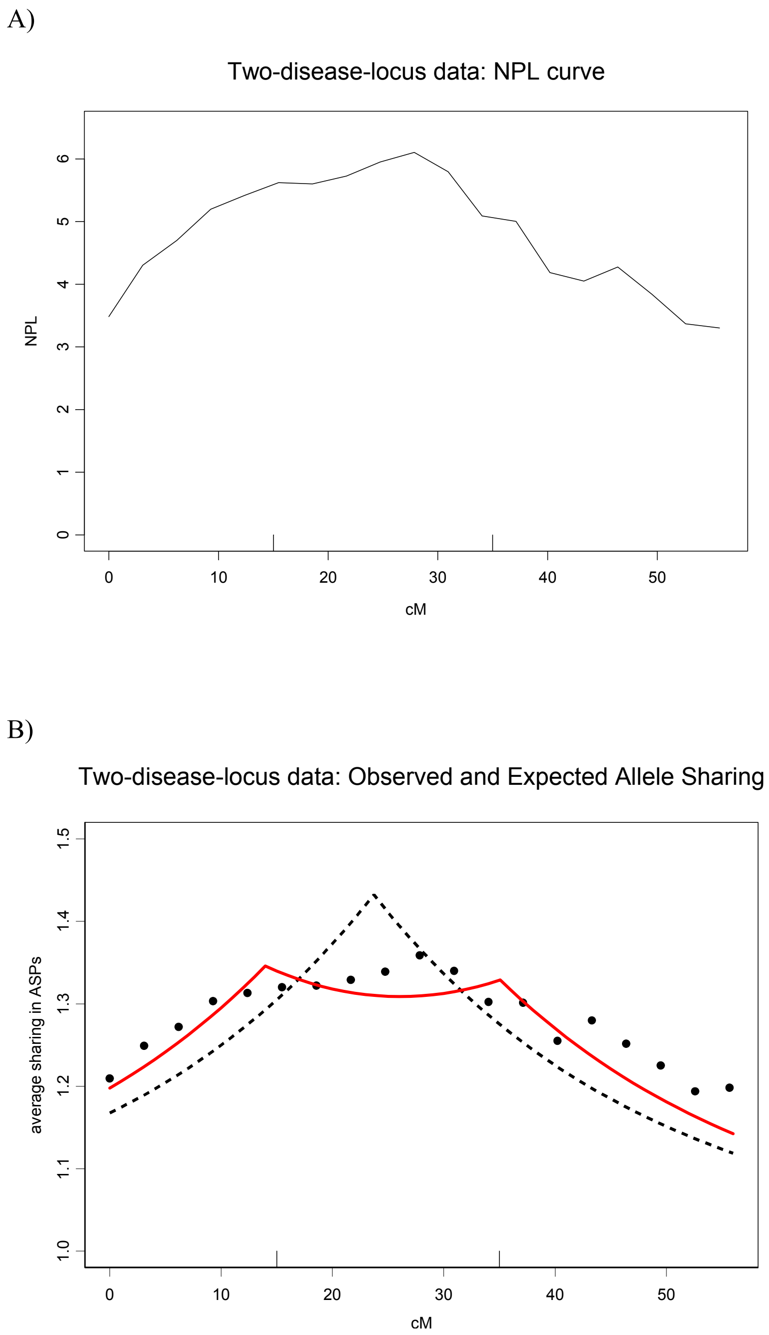


Figure 5.2: Plots for simulated example 2. (a) NPL plot; (b) observed average IBD allele-sharing in the sample of ASPs at the markers (points) with one-locus model fit (dashed line) and two-locus model fit (solid line). True disease gene locations are shown along the x-axis. The generating model under two linked disease loci is described in the text.

5.2 Monte Carlo Simulation Study Methods

5.2.1 Simulation study design

More extensive simulations were carried out to assess properties of the proposed estimation method including bias, standard errors, and confidence interval coverage of the parameter estimates. Simulation settings were chosen to study the effects of sample size, map properties such as number of markers and intermarker distances, and various aspects of the underlying genetic model including the magnitude of the *C* parameters, distance between the two disease genes, and correlation in IBD sharing at the two disease genes. Here we present the results of several sets of simulations which illustrate the main properties of the estimation method. Each simulation is based on 1000 replicates.

For some simulations (section 5.3.4) genotype data were generated using Allegro in order to assess properties of estimates resulting when data are not fully informative. However, for most of the simulations we generated fully informative IBD sharing for ASPs under a variety of two-locus models using our own data generation program. The joint IBD sharing distribution at the two disease genes is required to generate data under a given two-locus model. Therefore, in the next section, we explain calculation of the joint IBD sharing probabilities for a two-locus model. We then describe the generation of fully informative IBD sharing data in section 5.2.3.

5.2.2 Relationship of joint IBD allele sharing at two linked disease genes to the underlying genetic model

Given a two-locus disease generating model (a 3x3 penetrance matrix, and disease allele frequencies at two diallelic disease genes) and the distance between the two disease loci, the conditional IBD allele sharing distribution (i.e. $\Pr(S(\tau_1)=i, S(\tau_2)=j \mid \Phi)$ for $i,j=0,1,2$, where Φ denotes the affected status of the two sibs) can be calculated for a randomly mating population based on the following:

$$\Pr\left(S(\tau_1)=i, S(\tau_2)=j \mid \Phi\right)$$

$$=\frac{\Pr\left(S(\tau_1)=i, S(\tau_2)=j, \Phi\right)}{\Pr\left(\Phi\right)}$$

$$=\frac{\sum_{g_1,g_2}\Pr\left(S(\tau_1)=i, S(\tau_2)=j, g_1, g_2, \Phi\right)}{\sum_{i,j}\Pr\left(S(\tau_1)=i, S(\tau_2)=j, \Phi\right)}$$

where g_1 and g_2 are the genotypes of both sibs at the first locus and second locus, respectively,

$$=\frac{\sum_{g_1,g_2}\Pr\left(\Phi \mid S(\tau_1)=i, S(\tau_2)=j, g_1, g_2\right)\Pr\left(S(\tau_1)=i, S(\tau_2)=j, g_1, g_2\right)}{\sum_{i,j}\Pr\left(S(\tau_1)=i, S(\tau_2)=j, \Phi\right)}$$

$$=\frac{\sum_{g_1,g_2}\Pr\left(\Phi \mid g_1, g_2\right)\Pr\left(S(\tau_1)=i, S(\tau_2)=j, g_1, g_2\right)}{\sum_{i,j}\Pr\left(S(\tau_1)=i, S(\tau_2)=j, \Phi\right)}$$

$$=\frac{\sum_{g_1,g_2}\Pr\left(\text{1st sib } \mathit{aff} \mid g_1, g_2\right)\Pr\left(\text{2nd sib } \mathit{aff} \mid g_1, g_2\right)\Pr\left(S(\tau_1)=i, S(\tau_2)=j, g_1, g_2\right)}{\sum_{i,j}\Pr\left(S(\tau_1)=i, S(\tau_2)=j, \Phi\right)}$$

$$= \frac{\sum_{g_1,g_2} \Pr\left(aff_1 \mid g_1,g_2\right)\Pr\left(aff_2 \mid g_1,g_2\right)\Pr\left(g_1,g_2 \mid S(\tau_1)=i, S(\tau_2)=j\right)\Pr\left(S(\tau_1)=i,S(\tau_2)=j\right)}{\sum_{i,j}\Pr\left(S(\tau_1)=i,S(\tau_2)=j,\Phi\right)}.$$

Therefore, denoting the IBD sharing at τ_1 by S_1 and IBD sharing at τ_2 by S_2:

$$= \frac{\sum_{g_1,g_2} \Pr\left(aff_1 \mid g_1,g_2\right)\Pr\left(aff_2 \mid g_1,g_2\right)\Pr\left(g_1 \mid S_1=i\right)\Pr\left(g_2 \mid S_2=j\right)\Pr\left(S_1=i \mid S_2=j\right)\Pr\left(S_2=j\right)}{\sum_{i,j}\sum_{g_1,g_2} \Pr\left(aff_1 \mid g_1,g_2\right)\Pr\left(aff_2 \mid g_1,g_2\right)\Pr\left(g_1 \mid S_1=i\right)\Pr\left(g_2 \mid S_2=j\right)\Pr\left(S_1=i \mid S_2=j\right)\Pr\left(S_2=j\right)}.$$

Note that we have assumed that $\Pr(\Phi|g_1,g_2) = \Pr(1^{st}$ sib affected$|g_1,g_2)\cdot\Pr(2^{nd}$ sib affected$|g_1,g_2)$. This means that conditional on the genotypes at the two disease genes, the probability of the first sib being affected is independent of the probability of the second sib being affected. This independence assumption reflects an underlying assumption that there are no shared environmental or other genetic effects contributing to the disease. Although this is a fairly restrictive assumption, it is only made in the determination of the joint IBD distribution at the two disease genes, which is then used to generate data under a given two-locus disease model. In other words, the estimation method presented in the remainder of this chapter does not require the assumption of this relatively simple two-locus disease model.

It is shown above that in order to compute the joint IBD sharing probabilities at the two linked loci we need:

1) $\Pr\left(aff_1 \mid g_1,g_2\right)$ and $\Pr\left(aff_2 \mid g_1,g_2\right)$

2) $\Pr\left(g_1 \mid S(\tau_1)=i\right)$ and $\Pr\left(g_2 \mid S(\tau_2)=j\right)$

3) $\Pr\left(S(\tau_1)=i \mid S(\tau_2)=j\right)$ and

4) $\Pr\left(S(\tau_2)=j\right)$.

The probabilities in the first component listed above, are the penetrances corresponding to the genotype of each of the sibs. Note that $\Pr\left(aff_1 \mid g_1, g_2\right)$, i.e. the probability that the first sib is affected given the genotypes of the two sibs at both genes, depends only on the 2-locus genotype of the first sib. Similarly $\Pr\left(aff_2 \mid g_1, g_2\right)$ is the penetrance of the two-locus genotype of the second sib. The second component, $\Pr\left(g_1 \mid S(\tau_1) = i\right)$, the probability of a particular set of genotypes at gene 1, for a pair of relatives, given their IBD sharing at gene 1, has been tabulated for biallelic genes (Thomson 1975). The tabulated values can be used to determine $\Pr\left(g_1 \mid S(\tau_1) = i\right)$ and $\Pr\left(g_2 \mid S(\tau_2) = j\right)$ for every two-locus genotype of the sib pair. The third component, Pr($S(\tau_1)$=i|$S(\tau_2)$=j), can be obtained from Table 3.7. Finally, $\Pr\left(S(\tau_2) = j\right)$ is the IBD sharing probability for the pair of relatives. For sib pairs $\Pr\left(S(\tau_2) = 0,1,2\right) = (1/4, 1/2, 1/4)$. Multiplying the components listed above and summing over all possible 2-locus genotype of the two sibs, (g_1,g_2), gives the joint probability of the two sibs are affected and share i alleles IBD at the first disease gene and j alleles IBD at the second disease gene. Dividing by the probability that both sibs are affected gives the conditional probability, Pr($S(\tau_1)$=i, $S(\tau_2)$=j | Φ), as required.

Expected IBD sharing for each of the two disease genes, and therefore values of the parameters C_1 and C_2 defined in chapter 4, can be calculated using the nine joint allele sharing probabilities (Pr($S(\tau_1)$=i, $S(\tau_2)$=j | Φ) for i,j=0,1,2). Let p_{ij} = Pr($S(\tau_1)$=i, $S(\tau_2)$=j | Φ) for i,j=0,1,2. Then

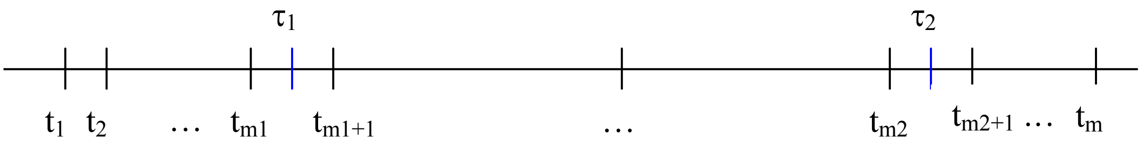
$$\mathrm{E}[S(\tau_1) \mid \Phi] = \sum_{j=0}^{2} p_{1j} + 2\sum_{j=0}^{2} p_{2j} \text{ and } \mathrm{E}[S(\tau_2) \mid \Phi] = \sum_{i=0}^{2} p_{i1} + 2\sum_{i=0}^{2} p_{i2} \ .$$

5.2.3 Generation of fully-informative IBD sharing data

A FORTRAN program was written to generate fully informative IBD sharing data for ASPs at any number of markers linked to 2 diallelic susceptibility genes. First, for a given underlying two-locus disease model (i.e. given a penetrance matrix for two diallelic loci, the disease allele frequencies, and the distance between the two disease genes), the joint IBD sharing probability distribution in a population of affected sib pairs was determined as described in section 5.2.2. Using this distribution, IBD sharing values at the two disease genes were generated for a sample of sib pairs. Then, for each sib pair in the sample, marker IBD sharing values were generated conditional on their IBD sharing at the two disease genes. The algorithm used to generate fully informative IBD sharing data for a map of markers is summarized below.

Algorithm used to generate fully informative IBD sharing data:

1. Given a disease generating model (2 locations of diallelic disease loci, a 3x3 penetrance matrix, and disease allele frequencies at the 2 disease loci), calculate the conditional allele sharing distribution (i.e. $\Pr(S(\tau_1)=i, S(\tau_2)=j \mid \Phi)$ for $i,j=0,1,2$, where Φ denotes the affected status of the 2 sibs) as described in section 5.2.2.
2. For each ASP, randomly generate the sharing status at the 2 disease loci, i.e. generate $(S(\tau_1), S(\tau_2))$ from the distribution $\Pr(S(\tau_1)=i, S(\tau_2)=j \mid \Phi)$.
3. Now assume there are m markers, $m1$ with $t_i<\tau_1$, and $m2$ with $t_i<\tau_2$. That is, assume the following map of markers:



and generate IBD sharing at the markers as follows:

3.1. Given the value of $S(\tau_1)$ generated in step 2, generate $S(t_{m1})$ from the conditional distribution $\Pr(S(t_{m1})|S(\tau_1))$. Then, from the conditional distribution $\Pr(S(t_{m1-1})|\ S(t_{m1}))$, generate $S(t_{m1-1})$ given the generated value of $S(t_{m1})$. Continue until $S(t_1)$ is generated.

3.2. Similarly, from the conditional distribution $\Pr(S(t_{m2+1})|S(\tau_2))$ generate $S(t_{m2+1})$. Then from the conditional distribution $\Pr(S(t_{m2+2})|\ S(t_{m2+1}))$ generate $S(t_{m2+2})$. Continue until $S(t_m)$ is generated.

3.3. Given the values of $S(\tau_1)$ and $S(\tau_2)$ generated in step 2, generate $S(t_{m1+1})$ from the conditional distribution $\Pr(S(t_{m1+1})|\ S(\tau_1),S(\tau_2))$. Then from the conditional distribution $\Pr(S(t_{m1+2})|\ S(t_{m1+1}),S(\tau_2))$ generate $S(t_{m1+2})$. Continue until $S(t_{m2})$ is generated.

Details:

In steps 3.1 and 3.2 the conditional IBD sharing distribution from Table 3.7 is used.

In step 3.3 note that for $t_i < t_k < t_j$, under the assumption of no interference:

$$\Pr\left(S(t_k)=a\mid S(t_i)=b,S(t_j)=c\right)=\frac{\Pr\left(S(t_j)=c\mid S(t_k)=a\right)\Pr\left(S(t_k)=a\mid S(t_i)=b\right)}{\Pr\left(S(t_j)=c\mid S(t_i)=b\right)}$$

Therefore the conditional probabilities from Table 3.7 can be used again (with the correct distances) to obtain the conditional probabilities of sharing a given number of alleles IBD at a locus, conditional on the sharing at two flanking loci.

5.3 Monte Carlo Simulation Study Results

5.3.1 Simulation study I: Different two-locus genetic models

In the first set of simulations we studied bias and standard deviations of estimates, magnitude of standard errors, and confidence interval coverage for four different underlying genetic models. The genetic models used in these simulations are described in Table 5.1. For this set of simulations, IBD sharing values for 500 independent ASPs were generated at 11 markers spaced at 10cM intervals (i.e. at 0,10,…,90,100cM).

Model	Penetrance matrix				Allele frequencies	Prevalence	τ_2-τ_1	C_1, C_2
A		aa	aA	AA	Pr(A) = .05	Approx 5%	30	.161,.161
	bb	0.050	0.050	0.950	Pr(B) = .05			
	bB	0.050	0.050	0.950				
	BB	0.950	0.950	0.950				
B		aa	aA	AA	Pr(A) = .015	Approx 1%	10	.327,.327
	bb	0.010	0.010	0.800	Pr(B) = .015		20	.283,.283
	bB	0.010	0.010	0.800			30	.254,.254
	BB	0.800	0.800	0.800			40	.235,.235
							50	.222,.222
C		aa	aA	AA	Pr(A) = 0.053	Approx 1%	30	.487,.487
	bb	0.000	0.000	0.000	Pr(B) = 0.053			
	bB	0.000	0.950	0.950				
	BB	0.000	0.950	0.950				
D		aa	aA	AA	Pr(A) = 0.02	Approx .5%	30	.348,.214
	bb	0.005	0.005	0.400	Pr(B) = 0.02			
	bB	0.005	0.005	0.400				
	BB	0.250	0.250	0.600				

Table 5.1: Models used for data generation for simulations presented in Tables 5.2-5.5, 5.7, and 5.8.

Average estimates of the four parameters in these simulations are very close to the true values (Tables 5.2), indicating that the estimates have little or no bias. Comparison of the average robust standard error estimates, obtained by the so-called sandwich estimator (Liang and Zeger 1986), to the empirical standard deviations computed across simulation replicates demonstrates that the robust standard error estimates for location estimates tend to be too low. Therefore, using these standard error estimates to construct confidence intervals (assuming an asymptotic normal distribution of the parameters), results in confidence intervals having lower than the nominal confidence level. However, confidence interval coverage reaches the nominal level for models with larger underlying C values (i.e. greater excess sharing at the two disease genes). Not surprisingly, performance is better (i.e. lower bias of estimates, smaller variances of estimates, higher CI coverage) when there is greater excess allele sharing in affected sib pairs at the two disease genes.

Confidence interval coverage is not only affected by the accuracy of the standard error estimates and bias of parameter estimates, but also by the distribution of the parameter estimators. Note that here, confidence intervals are constructed assuming the asymptotic normality of parameter estimates. Histograms of estimates of the four parameters $(C_1, C_2, \tau_1, \tau_2)$ for three simulations from Table 5.2 are shown in Figures 5.3-5.5. Although some of the histograms suggest departures from normality, in some situations it is difficult to draw conclusions about the distributions based on histograms. Several tests of goodness of fit were applied to test for departures from normality of the parameter estimates. Overall, these tests indicated that while estimates of C_1 and C_2 were approximately normal for some of the simulations, estimates of τ_1 and τ_2 showed evidence of lack of normality in all three of the simulations displayed in the histograms.

Model	Estimation of C_1 and C_2								Estimation of τ_1 and τ_2									
	Average Estimate		Mean bias		Empirical SD		Average robust SE		Average Estimate (cM)		Mean bias (cM)		Empirical SD		Average robust SE		"95%" CI coverage [a] (%)	
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
A	.163	.160	.003	-.001	.0377	.0401	.0389	.0404	34.8	65.8	-0.2	0.8	6.85	7.14	4.79	4.80	84.8	87.3
B	.255	.254	.001	.000	.0380	.0377	.0382	.0383	35.1	65.0	0.1	0.0	2.97	3.06	2.73	2.75	94.7	94.3
C	.486	.486	-.001	-.001	.0327	.0316	.0310	.0310	35.1	65.0	0.1	0.0	1.40	1.33	1.27	1.27	95.8	95.7
D	.345	.220	-.003	.006	.0399	.0462	.0383	.0402	34.9	65.1	-0.1	0.1	2.59	7.33	1.71	4.54	96.0	85.8

Table 5.2: Parameter estimation with different genetic models.

Simulation results are based on 1000 samples of 500 independent ASPs and a map of 11 markers spaced at 10 cM intervals (i.e. markers at 0,10,…,90,100 cM) with τ_1=35 and τ_2=65 (the resulting *C*'s can be found in the last column of Table 5.1).

[a] CI's were calculated as estimate +/- 1.96*se, using the robust se estimates.

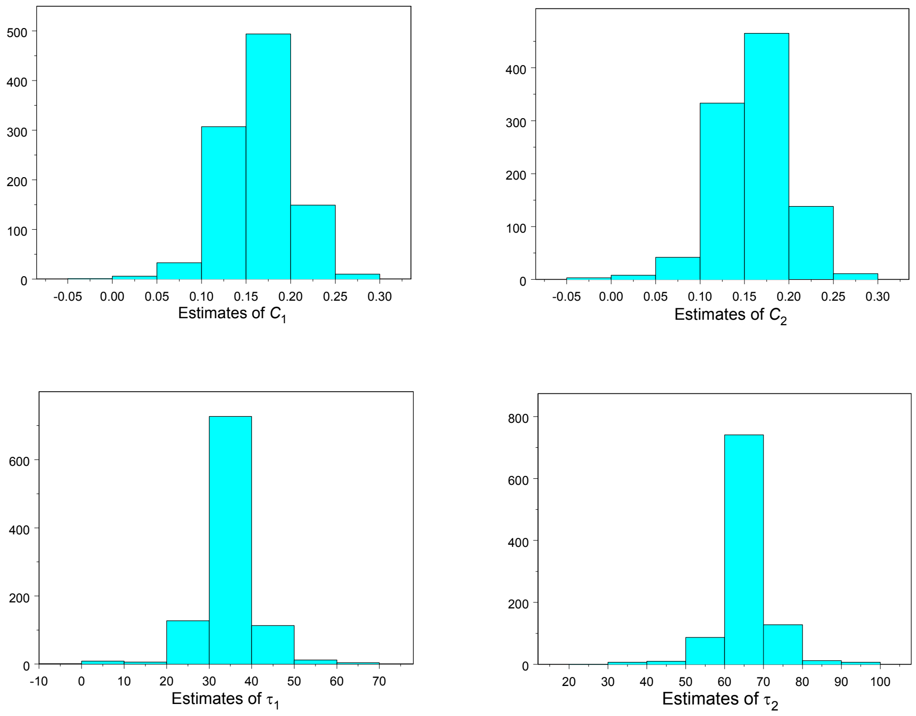


Figure 5.3: Histograms of parameter estimates under model A (for simulation presented in Table 5.2).

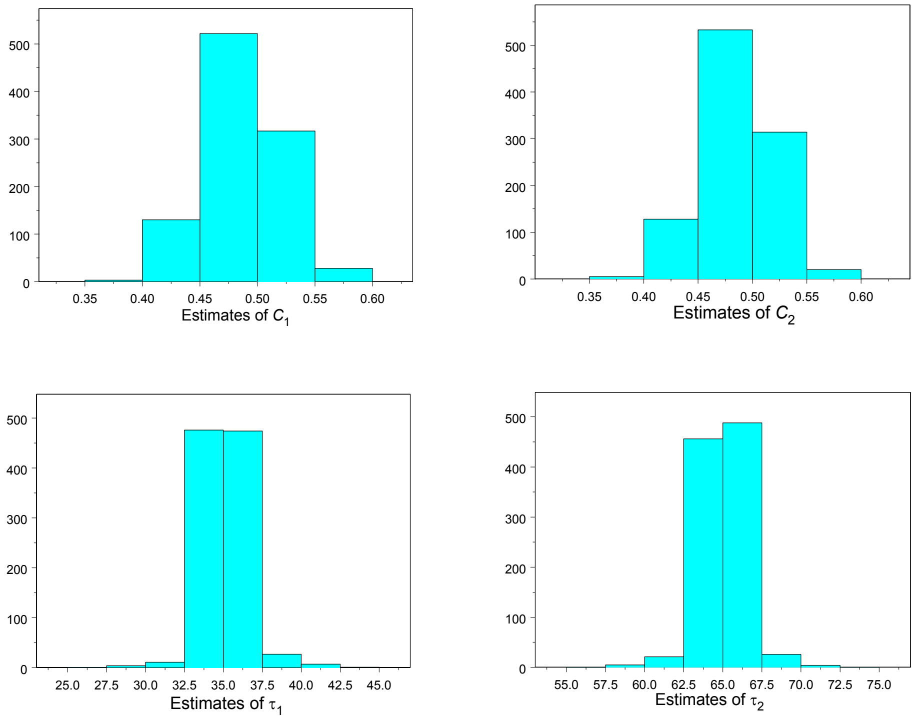


Figure 5.4: Histograms of parameter estimates under model B (for simulations presented in Table 5.2).

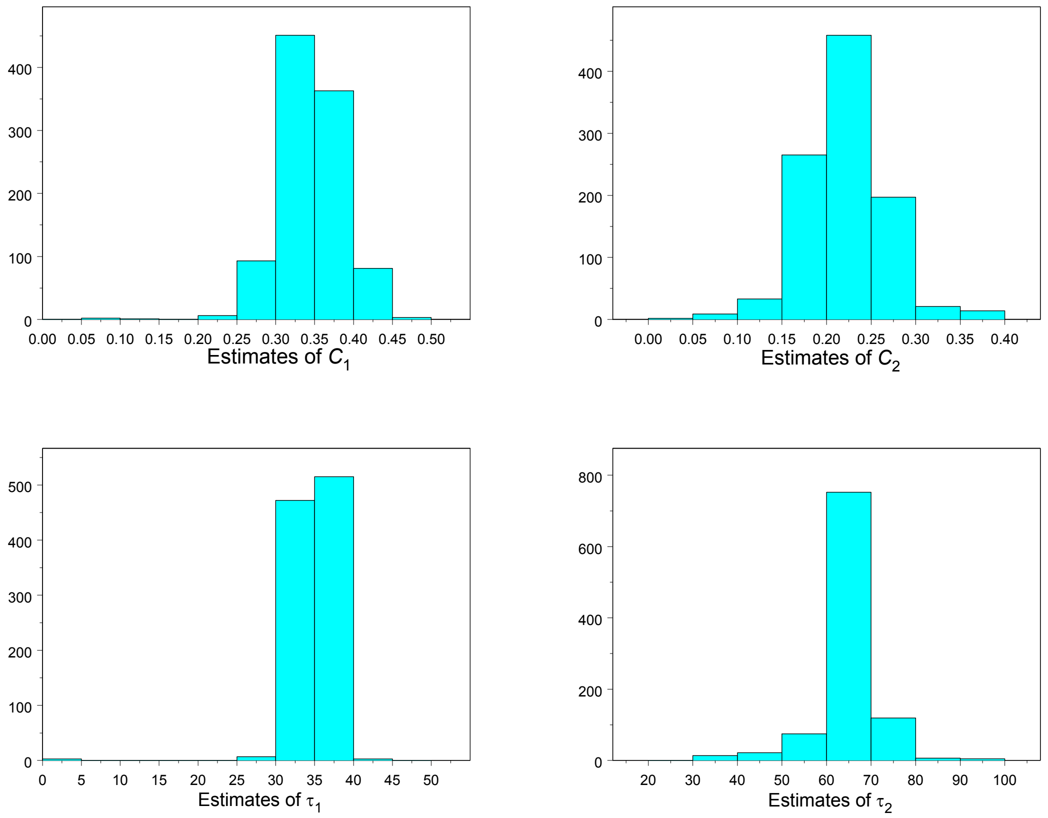


Figure 5.5: Histograms of parameter estimates under model D (for simulations presented in Table 5.2).

When the one-disease-locus model is fit to data generated under a model with two linked disease loci, on average, it results in an estimate of τ between the two true disease gene locations with the robust variance estimate substantially biased down, resulting in confidence intervals with low coverage of the two true locations (Table 5.3). It should be noted that when an incorrect model is fit to the data, solutions to the estimating equations are more dependent on the initial values provided for the estimation algorithm. In four of the simulations presented in Table 5.3, the initial value for τ was set to be the position mid-way between the two true locations, which led to estimates near that initial value and confidence intervals for τ with very low coverage of the two true disease gene locations. When the initial value for τ was set to be one of the two true locations, estimates of τ tended to be closer to that disease gene, and confidence intervals were more likely to include that one disease gene (rows two and five of Table 5.3). However, coverage of that one gene was still not sufficiently high, and there was still large bias in the estimate of its effect size, *C*.

Model	Initial	Estimation of *C*			Estimation of τ				
	τ value	Average estimate	SD_{emp}	SE_{rob}	Average estimate	SD_{emp} [a]	SE_{rob} [b]	"95%" CI coverage [c] (%)	
								τ_1	τ_2
A	50.0	.211	.0377	.0385	49.7	7.09	2.52	6.5	3.8
	35.0	.208	.0384	.0384	36.7	5.03	2.58	85.5	0.0
B	50.0	.334	.0396	.0384	49.9	5.37	1.63	0.3	0.1
C	50.0	.635	.0302	.0309	50.0	4.70	0.74	0.0	0.0
	35.0	.627	.0317	.0308	36.0	0.68	0.76	77.7	0.0
D	50.0	.382	.0386	.0384	41.6	3.09	1.43	19.0	0.0

Table 5.3: Results of fitting the one-locus model to data generated under a model with two linked disease genes.

Simulation results are based on 1000 samples of 500 independent ASPs and a map of 11 markers spaced at 10 cM intervals (i.e. markers at 0,10,…,90,100 cM) with τ_1=35 and τ_2=65 (the resulting *C*'s can be found in the last column of Table 5.1).

[a] SD_{emp} = empirical standard deviation of estimates for 1000 replicates in a simulation

[b] SE_{rob} = average robust standard error of estimates for 1000 replicates in a simulation

[c] CI coverage refers to the proportion of CIs for τ that cover τ_1 or τ_2.

5.3.2 Simulation study II: Effect of sample size and distance between loci

The second set of simulations was designed to demonstrate the effect of sample size and distance between the two disease genes on the estimates. Disease genes were placed 20, 30, or 50 cM apart. IBD sharing data at 11 markers spaced at 10cM intervals were generated for samples of 200, 500, and 1000 independent ASPs.

With a larger sample size there is less bias in estimates of the disease gene locations, and the variances of the location estimates are lower, leading to narrower confidence intervals with higher coverage (Table 5.4). Greater bias in location estimates and lower confidence interval coverage were observed when the two disease genes were close together.

True values		# ASPs	Average estimates		Mean bias		Empirical SD		Average robust SE		"95%" CI coverage (%)	
τ_1, τ_2	C_1, C_2		τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
40,60	.283,.283	200	39.6	60.4	-0.4	0.4	6.59	6.30	3.95	4.47	86.4	86.9
		500	39.7	60.4	-0.3	0.4	2.99	2.72	2.50	2.49	90.7	91.6
		1000	39.8	60.2	-0.2	0.2	1.88	1.76	1.81	1.81	92.1	93.7
35,65	.254,.254	200	35.0	65.4	0.0	0.4	6.99	7.00	4.93	4.64	87.8	86.5
		500	35.1	65.0	0.1	0.0	2.97	3.06	2.73	2.75	94.7	94.3
		1000	35.0	65.0	0.0	0.0	1.95	1.81	1.87	1.85	96.8	97.4
25,75	.222,.222	200	25.2	74.8	0.2	-0.2	6.10	6.63	4.53	4.75	89.0	87.9
		500	25.0	75.1	0.0	0.1	3.33	2.85	2.72	2.70	94.5	94.9
		1000	25.0	75.1	0.0	0.1	1.95	2.00	1.86	1.87	96.6	96.3

Table 5.4: Effect of sample size and distance between disease genes on estimation of τ_1 and τ_2. Data for these simulations was generated under model B described in Table 5.1. A map of 11 markers spaced at 10 cM intervals was used, i.e. markers at 0,10,…,90,100 cM.

5.3.3 Simulation study III: Effect of density and number of markers

The third set of simulations focused on how marker density and number of markers affect properties of estimates. Marker densities and total map distances were varied in these simulations (Table 5.5). For each simulation, data were generated for 500 independent ASPs under the same penetrance and allele frequency model (model B in Table 5.1). With this model, $C_1 = C_2 = 0.235$ if τ_1 and τ_2 are 40cM apart; $C_1 = C_2 = 0.283$ if they are 20 cM apart; and $C_1 = C_2 = 0.327$ if they are 10 cM apart.

A comparison of the first three rows of Table 5.5 indicates that while genotyping additional markers between the two disease loci may improve localization, having too many markers in the analysis can be detrimental in terms of confidence interval coverage. Further comparisons with other results presented in Table 5.5 establish that it is important to have several markers (3 or 4) between the two disease genes. However, genotyping additional markers far outside the region of the two disease genes may not be helpful and can decrease confidence interval coverage (see discussion in section 5.4).

τ_1, τ_2 (cM)	C_1, C_2	Distance between τ_1 and τ_2	# markers (total span of markers)	Marker spacing (cM)	Average estimate		Mean Bias		Empirical SD		Average robust SE		"95%" CI coverage (%)	
					τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
40,60	0.283,0.283	20	11 (0-100cM)	10	39.7	60.3	-0.3	0.3	2.61	2.61	2.44	2.45	92.0	91.1
			21 (0-100cM)	5	39.9	60.3	-0.1	0.3	2.62	2.47	2.32	2.35	94.2	94.0
			41 (0-100cM)	2.5	39.9	60.3	-0.1	0.3	2.84	2.65	2.20	2.22	88.5	88.3
30,70	0.235,0.235	40	11 (0-100cM)	10	29.8	70.4	-0.2	0.4	2.51	2.58	2.72	2.65	94.1	93.6
			21 (0-100cM)	5	29.8	70.0	-0.2	0.0	2.52	2.46	2.37	2.37	93.0	94.3
			41 (0-100cM)	2.5	29.8	70.0	-0.2	0.0	2.61	2.57	2.30	2.27	89.9	90.1
15,35	0.283,0.283	20	11 (0-50cM)	5	14.8	35.0	-0.2	0.0	1.81	1.98	1.90	1.90	95.0	93.8
6,16	0.327,0.327	10	12 (0-22cM)	2	5.9	16.0	-0.1	0.0	1.40	1.42	1.24	1.25	91.6	91.9

Table 5.5: Estimation of τ_1 and τ_2 with different marker maps (marker densities, number of markers). Based on 1000 replicates of 500 ASPs. Data generated under model B from Table 5.1.

5.3.4 Simulation study IV: Effect of marker informativeness

The fourth set of simulations compared the estimation procedure for data with different levels of marker informativeness. For these simulations, genotype data were generated assuming a heterogeneity model (as in simulated example 2) with 250 ASPs from each of two populations, for a total sample of 500 ASPs in each replicate. For both populations, genotypes at the same 11 markers were generated for ASPs and their parents using Allegro (Gudbjartsson et al. 2000). In the first population there was a single disease gene at position $\tau = 35.0$cM with expected IBD sharing of 1.325 in ASPs at this disease gene. In the second population there was a single disease gene at position $\tau = 65.0$ cM with expected IBD sharing of 1.325. Thus, in a population of affected sib pairs consisting of equal numbers of ASPs from each of the two populations described above, there are two disease susceptibility loci at 35 and 65 cM, with $C_1 = C_2 = 0.211$. IBD sharing was estimated from genotype data using GENEHUNTER (Kruglyak et al. 1996). Different levels of information content were attained by simulating genotypes at markers with 2, 3, 4, or 10 equally frequent alleles.

When our estimation method is applied to non-fully informative IBD sharing data with different levels of information content (Table 5.6), less informative markers lead to greater downward bias in estimates of the *C* parameters (i.e. lower estimates of expected sharing in ASPs at the disease genes). Furthermore, lower information content leads to greater bias in location estimates (two disease loci are estimated to be further apart than they really are) and thus to lower coverage of location confidence intervals.

# of	Approx	Estimation of C_1 and C_2				Estimation of τ_1 and τ_2									
alleles	info	Average		Mean Bias		Average		Mean Bias		Estimated SDs of estimates				"95%" CI	
	content[a]	estimate				estimate				Empirical		Robust		coverage (%)	
		C_1	C_2	C_1	C_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
10	92-95%	.210	.209	-.001	-.002	34.9	65.0	-0.1	0.0	3.65	4.09	3.22	3.40	93.7	91.8
4	80-86%	.200	.202	-.011	-.009	34.8	65.0	-0.2	0.0	3.93	3.69	3.11	3.09	90.5	92.1
3	72-79%	.192	.194	-.019	-.017	34.7	65.1	-0.3	0.1	3.59	3.78	3.06	3.04	92.9	91.4
2	52-62%	.170	.171	-.041	-.040	34.6	65.5	-0.4	0.5	3.95	3.75	2.86	2.84	89.9	90.8

Table 5.6: Evaluation of our estimation method applied to data from non-fully-informative markers.

Based on 1000 replicates of 500 ASPs (250 from population 1 + 250 from population 2, as described in the text). Different levels of information content (2rd column) are achieved by varying the number of equally frequent alleles at each marker (1st column). For all simulations in this table. $C_1 = C_2 = 0.211$, $\tau_1 = 35$, $\tau_2 = 65$.

[a] Approximate information content across the chromosome was obtained from GENEHUNTER 2.0 and is based on the definition of information content described by Kruglyak et al. (1996).

5.3.5 Simulation study V: Performance of the GEE method with an alternative parameterization

In section 4.1.2, an alternative parameterization of the two-locus model was described. With this alternative formulation of the mean function, a GEE approach can be used to estimate the parameters $\delta_2^*=(C_1^*,C_2^*,\tau_1,\tau_2)$. Results of several simulations using GEEs with this alternative parameterization of the two-locus model are shown in Table 5.7. Comparison of these results with those in Table 5.2 indicates that, in terms of disease gene localization, this parameterization provides overall very similar results to the original parameterization.

Model	Estimation of C_1^* and C_2^*										Estimation of τ_1 and τ_2									
	True Values		Average Estimate		Mean bias		Empirical SD		Average robust SE		Average Estimate (cM)		Mean bias (cM)		Empirical SD		Average robust SE		"95%" CI coverage [a] (%)	
	C_1^*	C_2^*	C_1^*	C_2^*	C_1^*	C_2^*	C_1^*	C_2^*	C_1^*	C_2^*	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
A	.124	.124	.125	.123	.001	-.001	.052	.053	.048	.048	34.9	65.6	-0.1	0.6	6.70	6.89	4.56	4.61	84.7	86.8
B	.195	.195	.196	.194	.001	-.001	.041	.041	.042	.042	35.1	65.0	0.1	0.0	2.99	3.07	2.73	2.75	94.4	94.1
C	.374	.374	.374	.373	.000	-.001	.037	.036	.035	.035	35.1	65.0	0.1	0.0	1.51	1.46	1.27	1.27	95.7	95.6
D	.311	.120	.304	.127	-.007	.007	.049	.045	.043	.043	34.7	64.9	-0.3	-0.1	2.79	6.26	1.72	4.53	95.1	87.8

Table 5.7: Estimation of $(C_1^*, C_2^*, \tau_1, \tau_2)$ for different genetic models.

See section 4.1.2 for definitions of the parameters C_1^* and C_2^*. Simulation results are based on 1000 samples of 500 independent ASPs and a map of 11 markers spaced at 10 cM intervals (i.e. markers at 0,10,…,90,100 cM) with τ_1=35 and τ_2=65. Data were generated under models A-D from Table 5.1.

[a] CI's were calculated as estimate +/- 1.96*se, using the robust (sandwich) se estimates.

5.4 Discussion

We have introduced and studied the properties of a method for simultaneous estimation of the locations of two linked disease susceptibility genes. With a sufficiently informative sample the proposed method provides unbiased estimates with confidence interval coverage near the nominal levels. The method produces better results when the two disease loci are further apart, when there is higher allele sharing at the two disease genes, and when more ASPs are available for analysis. When the two disease genes are believed to be located close to one another, a denser map of markers is recommended, to ensure that there are several markers (3-4) between the two disease genes.

When markers are not fully informative, estimates of the *C* parameters (expected IBD sharing among ASPs at the disease genes) are biased toward the null hypothesis of no gene effect. This is because when markers are not fully informative, and there really are disease gene(s) linked to the markers, then the estimated IBD sharing values (S_{im}^*) tend to be lower than the true unknown IBD sharing (S_{im}). This results in a downward bias in estimates of the *C*'s. The lower *C*'s are associated with greater bias in the disease gene location estimates, leading to lower coverage of the corresponding confidence intervals. How much lower the S_{im}^* values are compared to the S_{im} values, depends on the underlying genetic model (the true S_{im} values), and the level of information content available from the marker data.

In this study, confidence intervals were constructed using asymptotic GEE theory, under which the estimates are normally distributed and the robust variance estimator (also known as the sandwich estimator) provides consistent variance estimates. Although for some underlying genetic disease models with sufficient sample size, coverage of asymptotic theory confidence intervals reached the nominal levels, for some situations studied coverage was below the nominal level. Coverage of confidence intervals can be different from the nominal level for a number of reasons: biased parameter estimates (i.e. biased estimates of the τ's), biased robust variance estimates, or non-normality of the parameter estimates. Our simulation study indicated that the robust variance estimates of the location estimates tend to be downwardly biased. There was also evidence of lack of normality of the estimates. It is not unexpected for the robust estimator to yield downwardly biased variance estimates (for example see Pan 2001). Methods have been proposed to correct for the bias (e.g. Mancl and DeRouen 2001; Pan 2001) and the variability (e.g. Mancl and DeRouen 2001; Pan and Wall 2002) of the robust variance estimator. Such techniques could be applied here, in particular those that adjust for bias of the robust variance estimator. However, the improvements may be modest especially if assumptions of these methods are violated. Confidence interval coverage may be improved by replacing the robust variance estimates by bootstrap estimates of the variances. Confidence interval properties could potentially be further improved in situations when the estimates do not follow the asymptotic normal distribution, by estimating confidence interval endpoints by bootstrap quantiles.

Using a large number of markers in the analysis can lower the confidence interval coverage. Asymptotic properties of the GEE method are approached by increasing the number of independent sampling units in the analysis, which in this case is accomplished by increasing the number of independent affected sib pairs. By increasing the number of markers without increasing the number of ASPs, the clusters comprised of dependent data increase in size without proportionally increasing the effective sample size (i.e. number of independent units). This leads to a greater downward bias in the variance estimates provided by the robust variance estimator, and thus to less than nominal confidence interval coverage. This may also be exacerbated by the fact that we did not optimally model Cov($S_i^*|\Phi$). First of all, we assumed an independence correlation structure, which may decrease efficiency of parameter estimates (Liang and Zeger 1986). Furthermore, we did not model the variances Var($S_i^*(t_m)|\Phi$) (i.e. the diagonal elements of Cov($S_i^*|\Phi$)) in terms of a common set of parameters, making use of known relationships between genetic model parameters and variances of IBD sharing in the region. Rather, we empirically estimate the variance in IBD sharing at each marker, Var($S_i^*(t_m)|\Phi$) m=1,…,M, independent of the other markers. Essentially, this amounts to allowing a new parameter for the variance at each marker. As the number of markers increases, the number of unknown parameters estimated from the data increases, requiring a larger sample size to achieve asymptotic properties. Because we observed that using a large number of markers may lead to greater bias in the robust variance estimates and lower confidence interval coverage, we conclude that when a large number of markers are available for analysis, greater care must be taken in constructing the confidence intervals. Also, as the number of linked markers increases, modeling the variances and correlations in IBD sharing at these markers may become more beneficial.

Although confidence interval coverage can decrease as the number of markers increases, there are benefits resulting from the use of additional markers. Bias of parameter estimates may be reduced and the true variances of estimates, as estimated by the sample variances from all the replicates in a simulation, generally decrease. Thus, more accurate and more precise estimates can be achieved when more markers are used in the analysis. Greater benefits from using additional markers would be seen if the marker data were not fully informative.

We observed that fitting the one-locus model introduced by Liang et al. (2001*a*) can produce misleading results when there are two linked disease genes in the region. The single disease gene location is likely to be estimated at an incorrect position between the two true disease gene locations while the corresponding effect size tends to be over-estimated. Furthermore, the standard error of the location estimate tends to be biased down, leading to a confidence interval for a disease gene location that frequently does not cover either of the two true locations. This is not surprising, since the GEE approach is known to provide consistent estimates of the parameters under mild regularity conditions, provided that the mean function has been correctly specified, which is not the case if the one-locus mean function is postulated for data generated under a model with two linked disease loci. It should also be noted that the situation presented here, with the two loci having equal effect size, is a worst case scenario in terms of how different the estimates based on the two models will be. If one of the genes had a substantially higher effect than the other, as is the case with the chromosome 6 diabetes data application we present in chapter 8, the stronger gene may be localized with reasonable precision by fitting the one-locus model, although the estimate would be biased in the direction of the second gene.

When an incorrect (one-locus) model was assumed in the analysis, point and interval estimates of the parameters were quite dependent on the initial parameter values provided for the algorithm used in solving the GEEs. This problem persists in other situations, because the systems of equations are solved using a Newton-Raphson-type algorithm, which is known to be very sensitive to initial values. For example, the algorithm used to solve the two-locus GEE can also be quite sensitive to initial values when there is only one disease gene in the region or if at least one of the two genes has a small effect size. Simulation results demonstrating this sensitivity are shown in Table 5.8. This lack of robustness to initial value specification draws attention to the importance of careful selection of initial values. In the simulations, initial values are held fixed from replicate to replicate. In analysis of a single data set, more careful selection of initial values can be accomplished, for example by using plots or other forms of exploratory data analysis. Accuracy of variance estimates and confidence interval coverage may be improved by proper choice of initial values, although this is difficult to demonstrate in a simulation study.

Initial τ_1, τ_2	Estimation of C_1 and C_2		Estimation of τ_1 and τ_2						
	Average estimate		Average estimate		Empirical SD		Average Robust SE		"95%" CI coverage
	C_1	C_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	
50,60	.199	.132	48.4	63.9	9.59	12.79	3.02	4.99	625
50,70	.191	.099	46.6	73.2	12.15	16.81	5.03	13.63	830
40,60	.167	.167	42.7	57.8	15.08	14.62	5.60	4.52	398
30,70	.160	.158	39.2	61.6	17.92	17.95	7.22	6.94	558

Table 5.8: Effect of initial parameter values when fitting the two-locus model to data generated under a one-locus model with $C = 0.198$ and $\tau = 50$. Based on 1000 replicates of 500 ASPs. CI coverage is the number of replicates, out of 1000, in which at least one of the location confidence intervals included the true disease gene location. For some replicates confidence intervals could not be constructed either because no solution was obtained to the GEEs (convergence not attained) or the algorithm failed to estimate variances of the location parameters.

Although localization of two linked disease genes (i.e. estimation of τ_1 and τ_2) is the primary purpose of the method introduced in chapter 4, C_1 and C_2, i.e. the excess mean IBD sharing in ASPs at the two disease genes, are also estimated. In the one-locus model, mean excess IBD sharing at a single disease gene on a chromosome can be interpreted as the effect size of that gene. This effect size is 0 under the null hypothesis of no disease genes linked to the region. However, this simple "effect size" interpretation does not apply to C_1 and C_2 in the two-locus model. When two disease genes are linked, expected IBD sharing between affected relatives at any one of these genes increases as a result of effects of both genes. Therefore, for example, C_1 does not represent the effect of gene 1 alone, because even in the absence of any effect of gene 1, C_1 will increase as the distance between gene 1 and gene 2 decreases, due to the effect of gene 2. An alternative parameterization of the two-locus model, which simplifies the mean function formula, and is more conducive to a "gene effect size" interpretation of the parameters, is described in section 4.1.2. Several simulation results obtained by applying the GEE approach to estimate this alternative set of parameters in the two-locus model were presented in section 5.3.5, demonstrating that the two parameterizations provides very similar results. Both parameterizations have certain advantages in terms of interpretation.

Chapter 6

Tests for the Presence of One Disease Gene in a Region

In the preceding chapters we described GEE methods for fitting models with one or two disease-susceptibility genes in a single chromosomal region to ASP IBD sharing data. Simulations presented in chapter 5 revealed that fitting an incorrect model can lead to very biased estimates and confidence intervals with coverage below the nominal level. Thus, it is important to decide whether a one-locus or two-locus model is more appropriate. In chapter 7 we will discuss test statistics that would assist in making this decision. Before considering the more complex case of testing for two versus one disease genes in a region, in this chapter we propose several tests for one versus zero disease genes. In what follows, we assume there is one genotyped ASP per family, so that data for *n* independent ASPs are available.

6.1 Test Statistics for One versus Zero Disease Genes in a Region

In this chapter we consider the problem of testing the null hypothesis of no linkage to the region of interest versus the alternative of one disease gene in the region. Assume we obtain estimates of $\delta_1 = (C,\tau)$ using the GEE approach proposed by Liang et al. (2001*a*), that is by solving the equations:

$$\sum_{i=1}^{n}\left(\frac{\partial\mu_1(\delta_1)}{\partial\delta_1}\right)'\mathrm{Cov}_1^{-1}(S_i^* \mid \Phi_i)(S_i^* - \mu_1(\delta_1)) = 0$$

where $S_i^* = ((S_i^*(t_1),\ldots, S_i^*(t_M))'$, $\mu_1(\delta_1) = (\mu_1(t_1;\delta_1),\ldots, \mu_1(t_M;\delta_1))'$, and

$\mu_1(t;\delta_1) = E(S(t) \mid \Phi) = 1 + e^{-.04|t-\tau|}C$. Under the null hypothesis of no excess IBD sharing at locus τ, $C = 0$ and consequently $\mu_0(t) = 1$. We also obtain the null mean function, $\mu_0(t) = 1$, when $\tau = \infty$ (i.e. the disease gene is not linked to locus t). However, because τ was defined as the location of a disease gene within the map of markers, in the estimation procedure τ was restricted to lie in the marker map, i.e. $t_1 \leq \tau \leq t_M$. Thus, the null hypothesis $C = 0$ will be tested.

6.1.1 The "L-statistic" introduced by Liang et al. (2001*a*)

Let

$$L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \mathbf{1}'\mathrm{Cov}_0^{-1}\left(S_i^* \mid \Phi\right)\left(S_i^* - \mathbf{1}\right) \qquad (6.1)$$

where $\mathbf{1}=(1,\ldots,1)^T$ is a $M\times 1$ vector. Under the null hypothesis of no linkage:

$$\mathrm{E}_0\left(L\right) = 0 \quad \text{and} \quad \mathrm{Var}_0\left(L\right) = \sum_{i=1}^{n} \mathbf{1}'\mathrm{Cov}_0^{-1}\left(S_i^* \mid \Phi\right)\mathbf{1}\,.$$

Therefore, Liang et al. (2001*a*) proposed testing the null hypothesis of no linkage $C = 0$ versus the one-sided alternative hypothesis $C > 0$ using the statistic

$$L^* = \frac{L}{\left(\sum_{i=1}^{n} \mathbf{1}'\mathrm{Cov}_0^{-1}\left(S_i^* \mid \Phi\right)\mathbf{1}\right)^{1/2}} \tag{6.2}$$

which asymptotically has a standard normal distribution under H_0.

$\mathrm{Cov}_0(S_i^*)$ is the covariance matrix for the estimated IBD sharing at all the markers for the *i*th ASP, under the null hypothesis of no linkage. If the ASPs are all genotyped at the same set of markers then $\mathrm{Cov}_0(S_i^*)$ is identical for all ASPs, thus we can drop the "*i*" notation. Covariance of IBD sharing of fully informative linked markers is known, and does not depend on any unknown parameters. When markers are highly informative, $\mathrm{Cov}_0(S^*)$ can be approximated by $\mathrm{Cov}_0(S)$. Under the null hypothesis of no disease genes linked to the region,

$$\mathrm{Var}_0\left(S(t_j)\right) = 0.5 \text{ and}$$

$$\mathrm{Cov}_0\left(S(t_i), S(t_j)\right) = 0.5e^{-.04|t_i - t_j|}$$

i.e.

$$\mathrm{Cov}_0\left(S\right) = 0.5\begin{bmatrix} 1 & e^{-.04|t_2-t_1|} & e^{-.04|t_3-t_1|} & \cdots & e^{-.04|t_M-t_1|} \\ e^{-.04|t_1-t_2|} & 1 & e^{-.04|t_3-t_2|} & \cdots & e^{-.04|t_M-t_2|} \\ e^{-.04|t_1-t_3|} & e^{-.04|t_2-t_3|} & 1 & \cdots & e^{-.04|t_M-t_3|} \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ e^{-.04|t_1-t_M|} & e^{-.04|t_2-t_M|} & e^{-.04|t_3-t_M|} & \cdots & 1 \end{bmatrix} \tag{6.3}$$

One advantage of the statistic L^* is that it can be calculated using only the data and $\mathrm{Cov}_0(S)$, i.e. it does not require fitting the one-locus model.

6.1.2 Wald test

Under the null hypothesis of no linkage, when $C = 0$, the location of the disease gene is undefined, and the model does not depend on the parameter τ. Thus a test of the hypothesis $C = 0$ represents a reduction of the model which involves the removal of two parameters. Therefore, a two degree-of-freedom chi-squared test of the null hypothesis of no linkage vs. the alternative of linkage at the estimated disease gene location, $\hat{\tau}$, could be evaluated by testing H$_0$: $C = 0$ at $\tau = \tau_0$ vs. H$_A$: $C \neq 0$ using the test statistic

$$T_{wald} = \left[\begin{pmatrix} \hat{C} \\ \hat{\tau} \end{pmatrix} - \begin{pmatrix} 0 \\ \tau_0 \end{pmatrix}\right]^T \mathbf{C\hat{o}v}_R^{-1}(\hat{\delta}_1) \left[\begin{pmatrix} \hat{C} \\ \hat{\tau} \end{pmatrix} - \begin{pmatrix} 0 \\ \tau_0 \end{pmatrix}\right]$$

where $\mathbf{C\hat{o}v}_R(\hat{\delta}_1)$ is the robust estimate of the parameter covariance matrix. Since the estimated disease gene location is $\hat{\tau}$, we are interested in testing for linkage at $\tau_0 = \hat{\tau}$, leading to:

$$T_{wald} =$$

$$\begin{bmatrix} \hat{C} \\ 0 \end{bmatrix}^T \left[\left(\sum_{i=1}^{K} \hat{\mathbf{D}}^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}} \right)^{-1} \left(\sum_{i=1}^{K} \hat{\mathbf{D}}^T \hat{\mathbf{V}}_i^{-1} (S_i^* - \hat{\mu}_1)(S_i^* - \hat{\mu}_1)^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}} \right) \left(\sum_{i=1}^{K} \hat{\mathbf{D}}^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}} \right)^{-1} \right]^{-1} \begin{bmatrix} \hat{C} \\ 0 \end{bmatrix} \quad (6.5)$$

A p-value for this statistic can be obtained from a chi-squared distribution with 2df.

6.1.3 Approximate quasi-likelihood ratio tests

The two approximations to quasi-likelihood ratio tests described by McLeish and Small (1992) and Li (1993) can be applied, with

$\hat{\mu}_0(t) = 1$ (the mean function under the null hypothesis) and

$\hat{\mu}_1(t) = 1 + e^{-.04|t-\hat{\tau}|}\hat{C}$.

These tests also require Cov(S_i^*) under the null and/or alternative hypotheses.

The covariance matrix for fully informative data under the null hypothesis of no disease genes linked to the region, Cov$_0$(S_i), was shown in section 6.1.2 (equation 6.3). Under the hypothesis of exactly one disease gene in the region (or linked to the region) under consideration,

$$\mathrm{Var}_1\left(S_i(t_j)\right) = 0.5 + e^{-.08|t_j-\tau|}\left(\mathrm{Var}\left(S(\tau)\right) - 0.5\right) \text{ and}$$

$$\mathrm{Cov}_1\left(S_i(t_i), S_i(t_j)\right) = \begin{cases} e^{-.04|t_i-t_j|}\mathrm{Var}\left(S(\tau)\right) & \tau \in \left[t_i, t_j\right] \\ 0.5e^{-.04|t_i-\tau|-.04|t_j-\tau|}\left(\mathrm{Var}\left(S(\tau)\right) - 0.5\right) + 0.5e^{-.04|t_i-t_j|} & \tau \notin \left[t_i, t_j\right] \end{cases}$$

Also, it can be shown that Var($S(\tau)$) = p_2+p_0-(p_2-p_0)^2 = C-C^2+2p_0 where p_i=Pr($S(\tau)$=i|Φ). In the Genefinder program, Liang et al. (2001*a*) approximated Var($S(\tau)$) by C-C^2. Alternatively, one can estimate the unknown p_0 and use the correct formula for Var($S(\tau)$) as was described in chapter 3. We propose to approximate p_0 using the estimated proportion of ASPs that share 0 alleles IBD at the marker closest to $\hat{\tau}$. Using this estimate, the variances and covariances of IBD sharing at all the markers under the one-locus model can be estimated, and the M×M $\hat{\mathrm{Cov}}_1(S_i^*)$ matrix can be inverted numerically. Let $\hat{\mu}_1 = \mu_1(\hat{C}, \hat{\tau})$ The two approximate quasi-likelihood ratio test statistics can be computed using equations 2.9 and 2.10 from chapter 2, as follows:

$$\mathrm{LRT_{MS}} = -2\log\left\{\prod_{i=1}^{n}\left[1 + \left\{\mathbf{1} - \hat{\mu}_1\right\}^T \hat{\mathrm{Cov}}_1^{-1}(S_i^*)\left\{S_i^* - \hat{\mu}_1\right\}\right]\right\} \qquad (6.6)$$

$$\mathrm{LRT_{Li}} = -2\left[\sum_{i=1}^{n}\frac{1}{2}\left(\mathbf{1} - \hat{\mu}_1\right)^T \hat{\mathrm{Cov}}_1^{-1}(S_i^*)\left(S_i^* - \hat{\mu}_1\right) + \frac{1}{2}\left(\mathbf{1} - \hat{\mu}_1\right)^T \mathrm{Cov}_0^{-1}(S_i^*)\left(S_i^* - \mathbf{1}\right)\right] \qquad (6.7)$$

6.1.4 Modified quasi-score test

Assuming the Haldane map function and only one ASP per family, the quasi-score function for the one-locus model is given by:

$$U_1(C,\tau)=\begin{pmatrix} U_{1\_1}(C,\tau) \\ U_{2\_1}(C,\tau) \end{pmatrix}=\begin{pmatrix} \sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\mu}}{\partial C}\right)' \mathrm{Cov}_1^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\boldsymbol{\mu}_1\right) \\ \sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\mu}}{\partial \tau}\right)' \mathrm{Cov}_1^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\boldsymbol{\mu}_1\right) \end{pmatrix}$$

$$=\begin{pmatrix} \sum_{i=1}^{n}\begin{bmatrix} \frac{\partial\mu(t_1;C,\tau)}{\partial C} \\ \frac{\partial\mu(t_2;C,\tau)}{\partial C} \\ \vdots \\ \frac{\partial\mu(t_M;C,\tau)}{\partial C} \end{bmatrix}' \mathrm{Cov}_1^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\boldsymbol{\mu}_1\right) \\ \sum_{i=1}^{n}\begin{bmatrix} \frac{\partial\mu(t_1;C,\tau)}{\partial \tau} \\ \frac{\partial\mu(t_2;C,\tau)}{\partial \tau} \\ \vdots \\ \frac{\partial\mu(t_M;C,\tau)}{\partial \tau} \end{bmatrix}' \mathrm{Cov}_1^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\boldsymbol{\mu}_1\right) \end{pmatrix}=\begin{pmatrix} \sum_{i=1}^{n}\begin{bmatrix} e^{-.04|t_1-\tau|} \\ e^{-.04|t_2-\tau|} \\ \vdots \\ e^{-.04|t_M-\tau|} \end{bmatrix}' \mathrm{Cov}_1^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\boldsymbol{\mu}_1\right) \\ \sum_{i=1}^{n}\begin{bmatrix} k_1Ce^{-.04|t_1-\tau|} \\ k_2Ce^{-.04|t_2-\tau|} \\ \vdots \\ k_MCe^{-.04|t_M-\tau|} \end{bmatrix}' \mathrm{Cov}_1^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\boldsymbol{\mu}_1\right) \end{pmatrix} \quad (6.8)$$

where k_m = -0.04 for $t_m<\tau$, and k_m = 0.04 for $t_m>\tau$, $\mathbf{S}_i^*=\left(S_{i1}^*,...,S_{iM}^*\right)'$, $\boldsymbol{\mu}_1=\left(\mu_1,...,\mu_M\right)'$, S_{im}^* is the estimated IBD sharing for the ith sibpair (i=1,…,n) at the mth marker (m=1,…,M), $\mu_m=1+e^{-.04|t_m-\tau|}C$ is the expected IBD sharing at the mth marker, t_m is the position of the mth marker ($t_1<t_2<\ldots<t_M$).

Consider the null hypothesis $C = 0$. In order to construct the score test statistic for testing this hypothesis, the score statistic should be computed at $(C,\tau) = (0,\tilde{\tau})$ where $\tilde{\tau}$ is the GEE estimate of τ under the null hypothesis $C = 0$. However, under the reduced model of no linkage, τ is no longer defined, and cannot be estimated. One solution is to replace $\tilde{\tau}$ by $\hat{\tau}$ from the full (one-locus) model. Note that we are now testing for linkage specifically at the estimated location of the disease gene, $\hat{\tau}$, i.e. we are testing H$_0$: $C = 0$, $\tau = \hat{\tau}$. Thus, under the null hypothesis, the resulting statistic is expected to be asymptotically distributed as a chi-squared random variable with 2 degrees of freedom. Note that expectation and covariance of the data, $\mathbf{S}^*$, under the one-locus model evaluated at the null parameter values simplify as follows:

$$\boldsymbol{\mu}_1\left(C,\tau\right)\Big|_{C=0,\tau=\hat{\tau}} = \mathbf{1} \text{ , and } \operatorname{Cov}_1(\mathbf{S})\Big|_{C=0,\tau=\hat{\tau}} = \operatorname{Cov}_0(\mathbf{S})\,.$$

Also $\left.\dfrac{\partial\boldsymbol{\mu}}{\partial\tau}\right|_{C=0,\tau=\hat{\tau}} = \mathbf{0}$.

Therefore, substituting the null parameter values into equation 6.8, under H$_0$:

$$U_1\left(0,\hat{\tau}\right) = \begin{pmatrix} \sum_{i=1}^{n}\left(e^{-.04|t_1-\hat{\tau}|} \quad \ldots \quad e^{-.04|t_M-\hat{\tau}|}\right)' \operatorname{C\hat{o}v}_0^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^* - \mathbf{1}\right) \\ 0 \end{pmatrix},$$

and a score test statistic can be calculated as

$$T_{S\_\hat{\tau}} = U_1\left(0,\hat{\tau}\right)^T \hat{\Sigma}^{-1} U_1\left(0,\hat{\tau}\right) \tag{6.9}$$

with Σ estimated using the formula provided by Rotnizky and Jewell (1990; see equation 2.7 in chapter 2) or Breslow (1990, equation 2.8 in chapter 2) which leads to the statistic:

$$T_{S\_\hat{\tau}} = \frac{\left[\sum_{i=1}^{n}\left(e^{-.04|t_1-\hat{\tau}|} \quad \ldots \quad e^{-.04|t_M-\hat{\tau}|}\right)' \operatorname{C\hat{o}v}_0^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^* - \mathbf{1}\right)\right]^2}{\sum_{i=1}^{n}\begin{bmatrix} e^{-.04|t_1-\hat{\tau}|} \\ \vdots \\ e^{-.04|t_M-\hat{\tau}|}\end{bmatrix}' \operatorname{C\hat{o}v}_0^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^* - \mathbf{1}\right)\left(\mathbf{S}_i^* - \mathbf{1}\right)' \operatorname{C\hat{o}v}_0^{-1}(\mathbf{S}_i^*)\begin{bmatrix} e^{-.04|t_1-\hat{\tau}|} \\ \vdots \\ e^{-.04|t_M-\hat{\tau}|}\end{bmatrix}}.$$

Alternatively, rather than replacing $\tilde{\tau}$ by $\hat{\tau}$, we could integrate the score vector over τ, so that $U_1(C=0)$ could be evaluated without specifying τ.

Recall that

$$U_1\left(0,\tau\right)=\begin{pmatrix}\sum_{i=1}^{n}\left(e^{-.04|t_1-\tau|} \quad \ldots \quad e^{-.04|t_M-\tau|}\right)'\,\hat{\mathrm{Cov}}_0^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_\mathbf{i}^*-\mathbf{1}\right)\\ 0\end{pmatrix}.$$

A score test statistic could be constructed as

$$\frac{\left[U_{1\_1}\left(C=0\right)\right]^2}{\mathrm{Var}\left[U_{1\_1}\left(C=0\right)\right]}.$$

Assuming a uniform prior distribution for τ under the null hypothesis:

$$U_{1\_1}\left(C=0\right)=\int_\tau U_{1\_1}\left(0,\tau\right)d\tau=\int_\tau\sum_{i=1}^{n}\left(e^{-.04|t_1-\tau|} \quad \ldots \quad e^{-.04|t_M-\tau|}\right)'\,\mathrm{Cov}_0^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\mathbf{1}\right)$$

$$=\sum_{i=1}^{n}\left[\int_\tau\left(e^{-.04|t_1-\tau|} \quad \ldots \quad e^{-.04|t_M-\tau|}\right)'\mathrm{d}\tau\right]\mathrm{Cov}_0^{-1}(\mathbf{S}_i)\left(\mathbf{S}_i^*-\mathbf{1}\right).$$

For $\tau < t_m$: $\int_{-\infty}^{t_m} e^{-.04|t_m-\tau|}d\tau = 1/.04$; and for $\tau > t_m$: $\int_{t_m}^{\infty} e^{-.04|t_m-\tau|}d\tau = 1/.04$.

Thus a score test could be based on:

$$U_{1\_1}\left(C=0\right)=\sum_{i=1}^{n}\left(1/.04 \quad \ldots \quad 1/.04\right)'\,\hat{\mathrm{Cov}}_0^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\mathbf{1}\right)=\frac{1}{.04}\sum_{i=1}^{n}\mathbf{1}'\,\hat{\mathrm{Cov}}_0^{-1}(\mathbf{S}_i^*)\left(\mathbf{S}_i^*-\mathbf{1}\right)$$

Note that $U_{1\_1}(C=0)$ is a constant multiple of the statistic L described in section 6.1.1 (equation 6.2). Therefore, with a proper choice of variance function, a test based on $U_{1\_1}(C=0)$ would be equivalent to the L-test proposed by Liang et al. (2001*a*) described in section 6.1.2. Note that since the statistic L^* tests the hypothesis $C = 0$ without specifying a particular value of the location, τ, L^{*2} is a 1-df chi-squared statistic, and it tests for linkage anywhere in the region, rather than linkage at the disease gene location estimate, $\hat{\tau}$.

6.2 Assessment of Tests by Simulation

Properties of the test statistics described in section 6.1 were studied by simulation. The null distributions of the test statistics were examined by analyzing data generated under the null hypothesis. Power of the proposed tests was compared for a range of alternative models and sample sizes. Although the simulations were not extensive enough to accurately determine critical values for the test statistics, they did provide results that allow a general comparison of the different tests.

6.2.1 Simulation study design

True type I error or power of a test statistic under a given model was estimated by the proportion of significant test statistics obtained in a simulation consisting of 1000 replicates. Note that, with 1000 replicates, if the true type I error is 5%, 95% of the simulations are expected to yield an estimated type I error between 3.65% and 6.35%. Similarly, if the true power is 50%, 95% of the power estimates are expected to be between 46.9% and 53.1%. Although accurate estimates of type I error or power would require larger simulations, the simulations presented here were primarily intended to provide a preliminary comparison of the alternative test statistics. A general understanding of the behaviour of each test statistic and a relative comparison of properties of the tests were the main focus of this simulation study, rather than determining the absolute size/power of the tests.

For each replicate, fully informative marker IBD sharing data were generated for 200 or 1000 ASPs under a given underlying genetic model. Each replicate data set was analyzed by the one-locus GEE method described by Liang et al. (2001*a*), and all test statistics described in section 6.1 were calculated. It has been noted that the estimation procedure can be quite sensitive to initial values provided for the estimation algorithm, particularly for the gene location parameters. The simulation study should resemble “real data analysis” as much as possible. In an analysis of a single data set, a researcher would carefully choose initial values by using exploratory data analysis, graphical methods, or prior information from earlier studies. Furthermore, estimation should be repeated with different initial values to assess the robustness of the estimates to changes in initial values. This is difficult to accomplish in a simulation setting. For our simulations, the position of the marker with highest average IBD sharing in the sample was chosen as the initial value for the location parameter. If the algorithm for fitting the one-locus model did not converge to a solution, test statistics requiring estimates from the one locus model were deemed non-significant.

6.2.2 Simulation study results

Table 6.1 summarizes results of several simulations under the null hypothesis of no disease genes in the region. Results indicate that while the statistic L^* can be conservative for large sample sizes, the approximate quasi-likelihood ratio tests and modified quasi-score test tend to be anti-conservative. Type I errors of $\mathrm{LRT}_{\mathrm{Li}}$ are closer to the nominal 5% than those of $\mathrm{LRT}_{\mathrm{MS}}$, and $\mathrm{LRT}_{\mathrm{Li}}$ has very similar type I error to the score test.

In order to study power of the tests for one versus zero disease genes in a region, data were generated under several one-locus genetic models (Table 6.2). Power comparisons of the test statistics using chi-squared (or standard normal for L^*) critical values are shown in Table 6.3. Although results from Table 6.3 may suggest that L^* has the lowest power, power of the likelihood ratio tests and score test would be lower if the critical values for these tests

Sample	Marker	Type I error (%)				
size	map	L^*	Wald Test	$\mathrm{LRT}_{\mathrm{MS}}$	$\mathrm{LRT}_{\mathrm{Li}}$	Score Test
(#ASPs)	(spacing)		T_{wald}			$T_{S\_\hat{\tau}}$
1000	0-90 (10)	3.0	5.8	6.6	6.4	6.5
	0-90 (5)	4.7	5.6	6.3	6.0	6.4
500	0-90 (10)	4.6	5.4	6.0	6.0	5.9
	0-90 (5)	4.8	5.1	7.1	6.8	6.8
200	0-90 (10)	4.3	4.7	6.8	6.4	6.3
	0-90 (5)	5.6	4.8	7.3	6.6	6.9

Table 6.1: Type I error of nominal 5% tests for one-versus-zero disease genes in a region, using asymptotic critical values. (L^* is a one-sided test using critical values from a standard normal distribution, all other tests are two-sided tests using critical values from on a chi-squared distribution with 2df)

were properly adjusted to achieve correct type I error. Approximate empirical critical values for the different test statistics are shown in Table 6.4. Power was recalculated using these critical values (Table 6.5). Although the critical values used to approximate the true power are not expected to be highly accurate as they were based on only 1000 replicates, they nevertheless allow an improved assessment of power. Overall, after adjusting the critical values, the five statistics (L^*, Wald test T_{wald}, $\mathrm{LRT}_{\mathrm{MS}}$, $\mathrm{LRT}_{\mathrm{Li}}$, and score test $T_{S\_\hat{\tau}}$) have similar power. The expected trends of increased power with increased sample size or disease gene effect (larger C) are observed for all the statistics. The approximated quasi-likelihood ratio test based on the method of Li (1993), i.e. $\mathrm{LRT}_{\mathrm{Li}}$, is appealing because of the simplicity of the calculations involved and because it has comparable power to the other statistics. The test proposed by Liang et al. (2001*a*), L^*, has the advantage of not requiring the one-locus model to be fit. However, more simulations are needed to determine whether any of the statistics have strong advantages in specific types of data.

Model	Penetrance vector	Pr(D)	*C*
1-A	(0.015, 0.055, 0.055)	0.35	0.06
1-B	(0.015, 0.062 ,0.062)	0.29	0.08
1-C	(0.015, 0.075 ,0.075)	0.20	0.12
1-D	(0.015, 0.122, 0.122)	0.12	0.2
1-E	(0.015, 0.220 ,0.220)	0.055	0.3

Table 6.2: One-locus models used for power study. For each of these models the prevalence of the disease is approximately 3.8%. Pr(D) = frequency of the disease predisposing gene.

Model	*C*	Sample	Marker	Power (%)				
		size	map	L^*	Wald	LRT_{MS}	LRT_{Li}	Score
		(#ASPs)	(spacing)		Test			Test
					(T_{wald})			$(T_{S\_\hat{\tau}})$
1-A	.06	200	0-90 (10)	16.2	14.5	16.5	17.0	16.7
		1000	0-90 (10)	47.6	53.6	57.8	58.3	58.3
1-B	.08	200	0-90 (10)	25.4	24.2	27.1	27.8	27.2
		1000	0-90 (10)	65.8	78.6	80.7	81.4	81.7
1-C	.12	200	0-90 (10)	44.5	48.3	50.6	52.2	51.2
		1000	0-90 (10)	93.5	99.0	99.3	99.3	99.3
1-D	.2	200	0-90 (10)	75.6	87.7	89.2	89.6	89.7
		1000	0-90 (10)	100	100	100	100	100
1-E	.3	200	0-90 (10)	97.4	99.9	100	100	100
		1000	0-90 (10)	100	100	100	100	100

Table 6.3: Power of nominal 5% tests for one-versus-zero disease genes in a region. Asymptotic critical values (from a standard normal distribution for L^* and a chi-square distribution with 2 df for other tests) were used to determine significance. The disease gene was located at 45 cM ($\tau = 45$).

Sample size (#ASPs)	Marker map (spacing)	Approximate critical value for a test with $\alpha = 5\%$				
		L^*	Wald Test (T_{wald})	LRT_{MS}	LRT_{Li}	Score Test ($T_{S\_\hat{\tau}}$)
1000	0-90 (10)	1.42	6.27	6.80	6.62	6.77
200	0-90 (10)	1.58	5.89	7.14	6.40	6.50

Table 6.4: Approximate critical values of 5% tests for one-versus-zero disease genes in a region, based on 1000 replicates under the null hypothesis.

Model	C	Sample size (#ASPs)	Marker map (spacing)	Power (%)				
				L^*	Wald Test (T_{wald})	LRT_{MS}	LRT_{Li}	Score Test ($T_{S\_\hat{\tau}}$)
1-A	.06	200	0-90 (10)	18.1	15.5	12.1	14.5	14.1
		1000	0-90 (10)	56.1	51.6	50.1	53.2	51.0
1-B	.08	200	0-90 (10)	26.9	24.8	18.7	24.4	23.7
		1000	0-90 (10)	74.3	76.6	75.9	77.9	77.1
1-C	.12	200	0-90 (10)	47.8	48.8	40.9	48.2	46.9
		1000	0-90 (10)	95.6	98.9	98.6	98.8	98.7
1-D	.2	200	0-90 (10)	77.0	88.1	84.4	87.9	88.1
		1000	0-90 (10)	100	100	100	100	100
1-E	.3	200	0-90 (10)	97.8	99.9	100	100	100
		1000	0-90 (10)	100	100	100	100	100

Table 6.5: Approximate power of 5% tests for one-versus-zero disease genes in a region. Using approximate critical values obtained by simulation under the null hypothesis (Table 6.4).

Chapter 7

Tests for the Presence of Two Linked Disease Genes in One Region

Having preliminary evidence for at least one disease gene in the region, and perhaps data suggesting the possible existence of a second disease gene in the same region, we may wish to test whether the two-locus model fits the data significantly better than the one-locus model. In this chapter we discuss tests for evaluating evidence for two linked disease genes. Again, to simplify the discussion, we assume there is one genotyped ASP per family, so that data for *n* independent ASPs are available.

7.1 Issues in Formulating Tests for the Number of Disease Genes in a Region

7.1.1 Multiple roots of quasi-score estimating equations

One issue that arises in this testing situation and may lead to poor performance of some test statistics is the existence of multiple-roots of the quasi-score equations. The estimating equations for the two-locus model can have multiple roots. The existence of multiple-roots is easier to demonstrate in the re-parameterized two-locus model, written in terms of $\delta_2^{*}=(C_1^{*},C_2^{*},\tau_1,\tau_2)$ as was proposed in section 4.1.2. With this parameterization, the quasi-score vector is given by:

$$\mathbf{U}_2^{*}\left(\delta_2^{*}\right)=\begin{pmatrix} \sum_{i=1}^{nfam}\frac{\partial\boldsymbol{\mu}_2^{*}}{\partial C_1^{*}}\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \\ \sum_{i=1}^{nfam}\frac{\partial\boldsymbol{\mu}_2^{*}}{\partial C_2^{*}}\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \\ \sum_{i=1}^{nfam}\frac{\partial\boldsymbol{\mu}_2^{*}}{\partial \tau_1}\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \\ \sum_{i=1}^{nfam}\frac{\partial\boldsymbol{\mu}_2^{*}}{\partial \tau_2}\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \end{pmatrix}$$

$$=\begin{pmatrix} \sum_{i=1}^{nfam}\left(e^{-.04|t_1-\tau_1|} \quad \dots \quad e^{-.04|t_M-\tau_1|}\right)\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \\ \sum_{i=1}^{nfam}\left(e^{-.04|t_1-\tau_2|} \quad \dots \quad e^{-.04|t_M-\tau_2|}\right)\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \\ \sum_{i=1}^{nfam}\left(k_{11}C_1^{*}\,e^{-.04|t_1-\tau_1|} \quad \dots \quad k_{1M}C_1^{*}e^{-.04|t_M-\tau_1|}\right)\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \\ \sum_{i=1}^{nfam}\left(k_{21}C_2^{*}\,e^{-.04|t_1-\tau_2|} \quad \dots \quad k_{2M}C_2^{*}e^{-.04|t_M-\tau_2|}\right)\mathrm{Cov}_2^{-1}(\mathbf{S}^{*})\left(\mathbf{S}_{\mathbf{i}}^{*}-\boldsymbol{\mu}_2^{*}\right) \end{pmatrix} \qquad (7.1)$$

where $k_{1m}=\begin{cases}-0.04 & \text{if } t_m<\tau_1 \\ 0.04 & \text{if } t_m>\tau_1\end{cases}$ and $k_{2m}=\begin{cases}-0.04 & \text{if } t_m<\tau_2 \\ 0.04 & \text{if } t_m>\tau_2\end{cases}$.

Let:

$$\left(C_1^* = \alpha\hat{C}, C_2^* = (1-\alpha)\hat{C}, \tau_1 = \hat{\tau}, \tau_2 = \hat{\tau}\right) \text{ for some constant } \alpha,\ 0<\alpha<1.$$

Note that $\mu_2^*(t) = 1 + C_1^* e^{-.04|t-\tau_1|} + C_2^* e^{-.04|t-\tau_2|}$ evaluated at

$\left(C_1^* = \alpha\hat{C}, C_2^* = (1-\alpha)\hat{C}, \tau_1 = \hat{\tau}, \tau_2 = \hat{\tau}\right)$ is:

$$\begin{aligned}
&\mu_2^*\left(C_1^* = \alpha\hat{C}, C_2^* = (1-\alpha)\hat{C}, \tau_1 = \hat{\tau}, \tau_2 = \hat{\tau}\right) \\
&= 1 + \alpha\hat{C}e^{-.04|t-\hat{\tau}|} + (1-\alpha)\hat{C}e^{-.04|t-\hat{\tau}|} \\
&= 1 + \hat{C}e^{-.04|t-\hat{\tau}|}
\end{aligned}$$

i.e., the one locus mean function for parameters $C = \hat{C}$ and $\tau = \hat{\tau}$. Similarly,

$$\mathrm{Cov}_2\left(S^*; C_1^* = \alpha\hat{C}, C_2^* = (1-\alpha)\hat{C}, \tau_1 = \hat{\tau}, \tau_2 = \hat{\tau}\right) = \mathrm{Cov}_1\left(S^*; C = \hat{C}, \tau = \hat{\tau}\right)$$

Now consider evaluating the score function (equation 7.1) at

$\left(\tau_1 = \hat{\tau}, \tau_2 = \hat{\tau}, C_1^* = \alpha\hat{C}, C_2^* = (1-\alpha)\hat{C}\right)$:

$$\mathbf{U}_2^*\left(\tau_1 = \hat{\tau}, \tau_2 = \hat{\tau}, C_1^* = \alpha\hat{C}, C_2^* = (1-\alpha)\hat{C}\right)$$

$$= \begin{pmatrix}
\sum_{i=1}^{nfam}\left(e^{-.04|t_1-\hat{\tau}|} \quad \ldots \quad e^{-.04|t_M-\hat{\tau}|}\right)\mathrm{Cov}_1^{-1}(\mathbf{S}^*)\left(\mathbf{S}_\mathbf{i}^* - \hat{\boldsymbol{\mu}}_1\right) \\
\sum_{i=1}^{nfam}\left(e^{-.04|t_1-\hat{\tau}|} \quad \ldots \quad e^{-.04|t_M-\hat{\tau}|}\right)\mathrm{Cov}_1^{-1}(\mathbf{S}^*)\left(\mathbf{S}_\mathbf{i}^* - \hat{\boldsymbol{\mu}}_1\right) \\
\sum_{i=1}^{nfam}\left(k_1\alpha\hat{C}e^{-.04|t_1-\hat{\tau}|} \quad \ldots \quad k_M\alpha\hat{C}e^{-.04|t_M-\hat{\tau}|}\right)\mathrm{Cov}_1^{-1}(\mathbf{S}^*)\left(\mathbf{S}_\mathbf{i}^* - \hat{\boldsymbol{\mu}}_1\right) \\
\sum_{i=1}^{nfam}\left(k_1(1-\alpha)\hat{C}e^{-.04|t_1-\hat{\tau}|} \quad \ldots \quad k_M(1-\alpha)\hat{C}e^{-.04|t_M-\hat{\tau}|}\right)\mathrm{Cov}_1^{-1}(\mathbf{S}^*)\left(\mathbf{S}_\mathbf{i}^* - \hat{\boldsymbol{\mu}}_1\right)
\end{pmatrix}$$

Note that the first two components of the above score function are identical to the first component of the one-locus model score function evaluated at $\left(\hat{C}, \hat{\tau}\right)$, which equals zero since $\left(\hat{C}, \hat{\tau}\right)$ is a solution to the one-locus model quasi-score equations. The third and fourth components are constant (α or (1-α)) multiples of the second component of the one-locus model quasi-score evaluated at the one-locus model solution $\left(\hat{C},\hat{\tau}\right)$ and are, therefore, also equal to zero. Thus, it has been shown, that in this re-parameterized model, any solution of the form $\left(\tau_1 = \hat{\tau}, \tau_2 = \hat{\tau}, C_1^* = \alpha\hat{C}, C_2^* = (1-\alpha)\hat{C}\right)$ solves the two-locus quasi-score equations. Thus, these equations have multiple roots. Although it is not as simple to verify, it can also be demonstrated that the score equations of the two-locus model parameterized in terms of δ_2 = (C_1,C_2, τ_1, τ_2) may have multiple roots.

It has been suggested (McCullagh 1991) that a quasi-score test may have undesirable properties when the quasi-score equations have multiple roots. On the other hand, Li (1993) suggested that the approximate quasi-likelihood ratio he introduced may be used to distinguish between multiple roots in such situations. Thus, the approximate quasi-likelihood ratio tests may have better properties in this testing problem. It should be pointed out that finding an "incorrect root" with the two-locus GEE estimation algorithm (i.e. the one locus solution in the case of two linked disease genes) can often be avoided by careful selection of initial parameter values.

7.1.2 Undefined nuisance parameter under the null hypothesis

In developing the score statistics for one versus zero disease genes, it was noted that under the null hypothesis of no disease genes in the region, specified as $C = 0$, the location of a disease gene within the map, τ, is undefined and consequently indeterminate. A similar problem is encountered in testing for two versus one disease genes in a region. If a second potential disease gene does not have an effect on IBD sharing in the region, then it is not a disease gene for the phenotype under study, and hence the location of this second postulated disease gene in the region does not exist. This type of problem, where a nuisance parameter only exists under the alternative hypothesis, was discussed by Davies (1977, 1987).

One impact of this problem is that conventional test statistics are not computable, as seen, for example, in section 6.1.3 (for the one-versus-zero score test) and 7.2.3 (for the two-versus-one score test). We propose to compute the score statistics at the value of the nuisance parameter estimated under the alternative hypothesis. Note, however, that this is no longer a conventional score test statistic of the simple null hypothesis of no linkage to the region of interest.

Due to the difficulties described above, the test statistics we propose may be considered to be somewhat *ad hoc*. Asymptotic distributions of these test statistics have been conjectured, rather than rigorously derived, and need to be evaluated empirically. Derivation of asymptotic distributions of the statistics presents a challenging theoretical problem generally beyond the scope of this thesis research.

7.1.3 Approximation of critical values or p-values by simulation

For several reasons discussed in this chapter, some of the proposed test statistics may not have simple null distributions. If a test statistic does not have a well defined distribution under the null hypothesis, critical values or p-values have to be approximated by generating a large number of data sets under the null hypothesis, calculating the test statistic for each one, and using the results to determine the critical value or p-value. Whereas in testing for one versus zero disease genes in a region the null model of no disease genes in the region is clearly defined and corresponds to one underlying genetic model, this is not the case when testing for two versus one disease genes in a region. For tests of two versus one disease genes, the null hypothesis "one disease gene in the region" does not specify a single underlying genetic model. Therefore in order to generate data under the null hypothesis, some assumptions must be made about the true null model. A logical approach is to assume that the null one-locus model is a model specified by estimates obtained by fitting the one-locus model to the data. However, to generate IBD sharing data, we require the IBD sharing distribution (i.e. p_0,p_1,p_2, where p_j=Pr($S(\tau)$=j|Φ) as defined in section 3.3) not just the mean excess IBD sharing (C). The method for estimating the IBD sharing proportions (p_0,p_1,p_2) described in section 3.3 is therefore useful for determining the null genetic model to use for data generation in this situation.

7.2 Test Statistics for Two versus One Disease Genes in a Region

Estimates of C and τ can be obtained by fitting the one-locus model of Liang et al. (2001*a*) given by:

$$\mu_1(C,\tau;t) = 1 + e^{-.04|t-\tau|}C\,.$$

Estimates of $\delta_2 = (C_1,C_2,\tau_1,\tau_2)$ or $\delta_2^* = (C_1^*,C_2^*,\tau_1,\tau_2)$ may be obtained by fitting the two-locus model derived in chapter 4, parameterized in terms of δ_2:

$$\mu_2(C_1,C_2,\tau_1,\tau_2;t) = \begin{cases} 1 + e^{-.04|t-\tau_1|}\cdot C_1 & \text{for } t < \tau_1 < \tau_2 \\ 1 + C_1 e^{-.04(t-\tau_1)}\dfrac{1-e^{-.08(\tau_2-t)}}{1-e^{-.08(\tau_2-\tau_1)}} + C_2 e^{-.04(\tau_2-t)}\dfrac{1-e^{-.08(t-\tau_1)}}{1-e^{-.08(\tau_2-\tau_1)}} & \text{for } \tau_1 < t < \tau_2 \\ 1 + e^{-.04|t-\tau_2|}\cdot C_2 & \text{for } \tau_1 < \tau_2 < t \end{cases}$$

or in terms of δ_2^*:

$$\mu_2^*(t) = 1 + C_1^* e^{-.04|t-\tau_1|} + C_2^* e^{-.04|t-\tau_2|} \qquad \text{for all } t.$$

In chapter 6 a number of tests for evaluating the evidence for a single disease gene in a region were described. These included Wald tests, approximate quasi-likelihood ratio tests, a score test, and a test based on the L^* statistic previously proposed by Liang et al. (2001*a*). In simulations presented in section 6.2, all of the test statistics showed similar performance in terms of type I error and power. The simplicity of the approximate likelihood ratio tests combined with their acceptable performance in comparison to the other tests makes them particularly appealing. Therefore, in developing tests for two-versus-one disease genes in a region, we focus on applying and evaluating the approximate quasi-likelihood ratio tests, in particular the test proposed by Li (1993). We also address other test statistics and issues associated with applying them to this testing problem.

7.2.1 Approximate quasi-likelihood ratio tests

Let

$$\hat{\mu}_1 = \left[\mu_1(\hat{C}, \hat{\tau}; t_1), \ldots, \mu_1(\hat{C}, \hat{\tau}; t_M) \right] \text{ and}$$

$$\hat{\mu}_2 = \left[\mu_2(\hat{C}_1, \hat{C}_2, \hat{\tau}_1, \hat{\tau}_2; t_1), \ldots, \mu_2(\hat{C}_1, \hat{C}_2, \hat{\tau}_1, \hat{\tau}_2; t_M) \right].$$

Also let Cov$_1$(S^*) be the variance/covariance matrix of the IBD sharing data under the null one-locus model, calculated as described in section 6.1.3. Cov$_2$(S^*) represents the covariance of the observations under the two-locus model. We empirically estimate Cov$_2$(S^*) using IBD sharing information from all the ASPs as was described in chapter 4 in the context of estimation of the two-locus model parameters by the GEE approach.

Recall that S^{*}_{ij} is the (estimated) IBD sharing at the *j*th marker for the *i*th sibpair, and

$\mathbf{S}^{*}_{\mathbf{i}} = \left(S^{*}_{i1}, S^{*}_{i2}, ..., S^{*}_{iM}\right)^{T}$ is the vector of IBD sharing at all the markers for the *i*th sibpair.

Following McLeish and Small (1992; equation 2.8 in chapter 2) and Li (1993; equation 2.9 in chapter 2), respectively, the two approximations to likelihood ratio statistics are calculated as:

$$\mathrm{LRT}_{\mathrm{MS}} = -2\log\left\{\prod_{i=1}^{nfam}\left[1+\left(\hat{\boldsymbol{\mu}}_{\mathbf{1}}-\hat{\boldsymbol{\mu}}_{\mathbf{2}}\right)^{\mathrm{T}}\,\mathrm{C\hat{o}v}_{2}^{-1}(\mathbf{S}^{*}_{\mathbf{i}})\left(\mathbf{S}^{*}_{\mathbf{i}}-\hat{\boldsymbol{\mu}}_{\mathbf{2}}\right)\right]\right\} \tag{7.2}$$

and

$$\begin{aligned}&\mathrm{LRT}_{\mathrm{Li}}\\&=-2\left[\sum_{i=1}^{nfam}\left\{\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_{\mathbf{1}}-\hat{\boldsymbol{\mu}}_{\mathbf{2}}\right)^{T}\mathrm{C\hat{o}v}_{2}^{-1}(\mathbf{S}^{*}_{\mathbf{i}})\left(\mathbf{S}^{*}_{\mathbf{i}}-\hat{\boldsymbol{\mu}}_{\mathbf{2}}\right)+\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_{\mathbf{1}}-\hat{\boldsymbol{\mu}}_{\mathbf{2}}\right)^{T}\mathrm{C\hat{o}v}_{1}^{-1}(\mathbf{S}^{*}_{\mathbf{i}})\left(\mathbf{S}^{*}_{\mathbf{i}}-\hat{\boldsymbol{\mu}}_{\mathbf{1}}\right)\right\}\right]\end{aligned} \tag{7.3}$$

Under the null hypothesis, the asymptotic distributions of these statistics are expected to approximate a chi-squared distribution with 2 degrees of freedom. However the distribution of these statistics may deviate from this theoretical distribution, particularly in small samples, for reasons including the fact that they are based only on approximations to quasi-likelihood ratios, as well as issues discussed in section 7.1. Therefore, properties of these tests were studied by simulation and will be discussed in section 7.3.

7.2.2 Wald Test

In order to test the null hypothesis of one disease gene in the region using a Wald test, we needed to state the null hypothesis as a set of restrictions on the parameters δ_2 or δ_2^{*}. Aitchison and Silvey (1960) discussed two ways of specifying restrictions: as constraint equations in the parameters, or as "freedom equations" expressing the parameters in terms of a second smaller set of parameters. In the problem of testing for the presence of two linked disease genes, the latter way of stating the null hypothesis arises more naturally, making the score test more easily applicable than the Wald test.

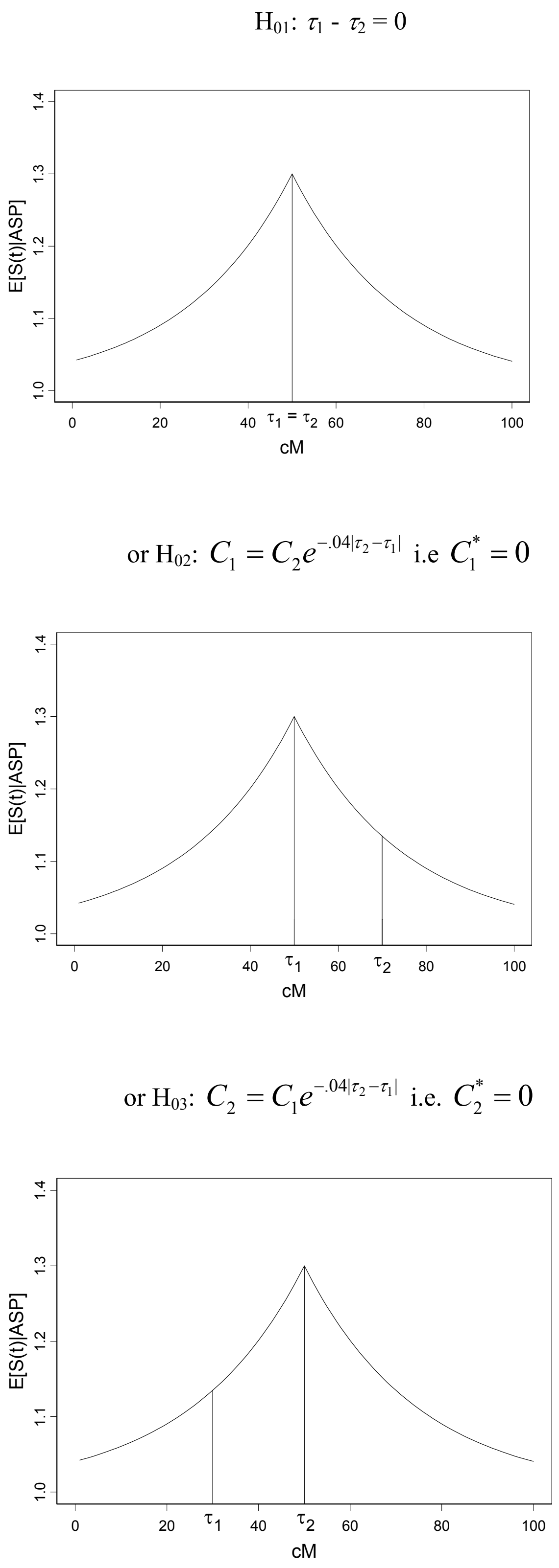
Because of certain complications, a test analogous to the Wald test presented in section 6.1.2 is not readily applicable to the problem of testing for two versus one disease genes in a region. Note that a two-locus mean function (parameterized by δ_2 or δ_2^*), simplifies to a one-locus mean function if the two disease gene loci are at the same location ($\tau_1 = \tau_2 = \tau$). If the two locations are different ($\tau_1 \neq \tau_2$), then a two-locus mean function simplifies to a one-locus mean function if one of the two loci (the "disease gene") increases IBD sharing in the region, while the other one does not. In the case of the δ_2^*-parameterized model, this means that $C_1^* = 0$ or $C_2^* = 0$. In the δ_2-parameterized model, if the gene at τ_1 is the only disease gene in the region then $C_2 = C_1 e^{-.04|\tau_2 - \tau_1|}$, and if the gene at τ_2 is the only disease gene in the region then $C_1 = C_2 e^{-.04|\tau_2 - \tau_1|}$. Therefore, in terms of the δ_2^*-parameterized model, the null hypothesis can be stated as

$$H_0\text{: } \tau_1 - \tau_2 = 0 \text{ or } C_1^* = 0 \text{ or } C_2^* = 0,$$

and in the δ_2-parameterized model, the null hypothesis can be stated as

$$H_0\text{: } \tau_1 - \tau_2 = 0 \text{ or } C_1 = C_2 e^{-.04|\tau_2 - \tau_1|} \text{ or } C_2 = C_1 e^{-.04|\tau_2 - \tau_1|}.$$

A graphical representation of this null hypothesis is shown in Figure 7.1.



H_{01}: $\tau_1 - \tau_2 = 0$

or H_{02}: $C_1 = C_2 e^{-.04|\tau_2 - \tau_1|}$ i.e $C_1^* = 0$

or H_{03}: $C_2 = C_1 e^{-.04|\tau_2 - \tau_1|}$ i.e. $C_2^* = 0$

Figure 7.1: Diagram of null hypothesis for the Wald test for evaluating evidence for two linked disease genes

Because in both cases the null hypothesis is a union of hypotheses, an intersection-union test can be applied, with the rejection region for the overall hypothesis being given by the intersection of the rejection regions for the three individual hypotheses. Three Wald statistics can be formulated for testing the three components of the null hypothesis. If all three Wald statistics exceed the appropriate critical values, there is evidence that there are two distinct loci ($\tau_1 \neq \tau_2$) both of which influence the IBD sharing in the region ($C_1^* \neq 0$ and $C_2^* \neq 0$) and the null hypothesis of one disease gene in the region is rejected. Thus we reject the null hypothesis when the minimum of the three statistics exceeds the critical value. The three component statistics for the δ_2-parameterized model are shown in Appendix C. Some weaknesses of the intersection-union test include the fact that, although an upper bound for the size of the test can be determined, the exact size may be difficult to determine, and frequently they have low power.

7.2.3 Modified quasi-score tests

It was mentioned in section 7.2.2, that the null hypothesis of one disease gene in the region is more naturally specified in terms of what has been termed “freedom equations” (Aitchison and Silvey 1959), i.e., via a reparameterization rather than via constraint equations. This statement of hypotheses arises more naturally because the null hypothesis is best described in terms of one set of parameters, $\delta_1 = (C, \tau)$, whereas the alternative is easiest to describe in terms of another set of parameters, $\delta_2 = (C_1, C_2, \tau_1, \tau_2)$. In terms of such equations, under the null hypothesis of one disease gene in the region, the excess IBD sharing at any locus should equal the mean IBD sharing given by the one-locus model, i.e. $E(S(t) \mid \Phi) - 1 = Ce^{-.04|t-\tau|}$. Therefore, in terms of the parameters $\delta_2 = (C_1, C_2, \tau_1, \tau_2)$, under the null one-locus model parameterized by $\delta = (C, \tau)$, $C_1 = Ce^{-.04|\tau_1 - \tau|}$ and $C_2 = Ce^{-.04|\tau_2 - \tau|}$, for any τ_1 and τ_2, and in particular for $\tau_1 = \hat{\tau}_1$ and $\tau_2 = \hat{\tau}_2$.

It was shown in section 6.1.4 that in order to calculate a score statistic to evaluate evidence for one disease gene in a region, the "null" value of the location parameter τ was needed. However, under the null hypothesis of no disease genes in the region, the location of a disease gene within the map, τ, is undefined and consequently indeterminate. A similar problem arises in developing a score test to evaluate evidence for the presence of two linked disease genes. In order to evaluate the two-locus model score function under the null hypothesis of one disease gene located at τ, a location of the second putative gene, say τ_2, has to be specified. We denote the null value of τ_2 by $\tilde{\tau}_2$. However, under the null hypothesis of one disease gene in the region (located at τ), the location of a second disease gene (τ_2) is not defined. There are several possible approaches that could be considered, including a solution parallel to that applied in section 6.1.4 to the testing for one versus zero disease genes.

One solution is to let $\tilde{\tau}_2 = \hat{\tau}_2$, the estimate of the second disease gene location under the two-locus model. The resulting statistic would test whether, after accounting for linkage to one disease gene in the region, there is further evidence of linkage to another disease gene at the locus $\hat{\tau}_2$. We could similarly construct the test statistic by evaluating the score at the other location identified by fitting the two-locus model, i.e., $\tilde{\tau}_2 = \hat{\tau}_1$. Thus two statistics can be calculated to evaluate evidence against the two hypotheses H$_{01}$: $C_1 = Ce^{-.04|\tau_1 - \tau|}$ at $\tau_1 = \hat{\tau}_1$ and H$_{02}$: $C_2 = Ce^{-.04|\tau_2 - \tau|}$ at $\tau_2 = \hat{\tau}_2$. These two statistics, $T_{S\_\hat{\tau}_1}$ and $T_{S\_\hat{\tau}_2}$, are shown in Appendix C. Under the null hypothesis of one disease gene in the region, there should be no additional effect on the IBD sharing of either locus ($\hat{\tau}_1$ or $\hat{\tau}_2$). Therefore, we reject the null hypothesis of one disease gene in the region if either $T_{S\_\hat{\tau}_1} > c$ or $T_{S\_\hat{\tau}_2} > c$, i.e. if

$$T_{S\_\max} = \max\left(T_{S\_\hat{\tau}_1}, T_{S\_\hat{\tau}_2}\right) > c$$

for an appropriately chosen critical value *c* to give a test with a pre-specified significance level. Note that this is a union-intersection test of the overall null hypothesis $C_1 = Ce^{-.04|\tau_1 - \tau|}$ and $C_2 = Ce^{-.04|\tau_2 - \tau|}$, for $\tau_1 = \hat{\tau}_1$ and $\tau_2 = \hat{\tau}_2$. It is generally difficult to derive the size of a union-intersection test. Although a lower bound for the size can be determined, this is not a useful result. In this situation, the critical value for a size α test can be determined by simulation. Alternatively, the two statistics could be added to get an overall test statistic, i.e. $T_{S\_\mathrm{sum}} = T_{S\_\hat{\tau}_1} + T_{S\_\hat{\tau}_2}$. Significance of this statistic would also be assessed by simulation.

7.3 Assessment of Tests by Simulation

As was mentioned at the beginning of this chapter, the approximate likelihood ratio test based on the method of Li (1993) is particularly appealing in the problem of testing for two versus one disease genes in one region. Issues associated with the alternative tests (based on Wald and score statistics) suggest that the null distributions of these statistics are likely to be difficult to determine. Therefore, critical values for these statistics would have to be determined by simulation. Some difficulties associated with obtaining such empirical critical values were noted in section 7.1.3. In section 7.3.1, we summarize results from a small simulation study comparing performance of the various tests for two-versus-one disease genes, while the more comprehensive simulation study presented in section 7.3.2 focuses on the performance of LRT_{Li}.

As in the simulations in chapter 6, true type I error and power under a given model were estimated by the proportion of significant test statistics obtained in a simulation consisting of 1000 replicates. This relatively small number of replicates will not provide highly accurate estimates of type I error or power, but will allow a general understanding of the behaviour of the statistics for a range of null and alternative models.

7.3.1 Comparison of properties of alternative test statistics

To compare type I errors of the various tests for two versus one disease genes in a region, data were generated under the null hypothesis of one disease gene for three different one-locus models (models C, D, and E from Table 6.2). Models A and B were not used for these simulations, since under these models there is generally very little evidence for even one disease gene in the region, and therefore it is unlikely that data arising under these models would be used to test for two disease genes in the region. We recommend using the methods for localizing and testing for two linked disease genes if the data, or previous studies, suggest the presence of one, and possibly two disease genes in one region. Thus we only apply the methods to models that would provide some indication of at least one gene. Observed type I errors of the test for two versus one disease genes in a region are shown in Table 7.1. Although the Wald and score test statistics are not expected to have chi-squared distributions with 2 df, critical values from a $\chi^2_{df=2}$ distribution were initially used for the Wald intersection-union test and the score test based on a maximum of the two score statistics evaluated at $\hat{\tau}_1$ and $\hat{\tau}_2$. This led to inflated type I errors for these statistics.

Therefore, to properly evaluate the power of these tests, empirical critical values need to be estimated. $\mathrm{LRT}_{\mathrm{Li}}$ has the strong advantage that the observed type I errors are close to the nominal 5%, and thus determining critical values for this statistic by simulation is not essential. Approximate critical values for the alternative test statistics based on the same set of one-locus models are shown in Table 7.2.

Two models used for comparing power of the different statistics are shown in Table 7.3, and power estimates are shown in Table 7.4. Results suggest that the Wald intersection-union test has the lowest power of the tests we considered. Score tests based on the maximum or sum of two score statistics evaluated at $\hat{\tau}_1$ and $\hat{\tau}_2$, have power comparable to that of the likelihood ratio tests. Under certain genetic models, $\mathrm{LRT}_{\mathrm{Li}}$ appears to be more powerful than $\mathrm{LRT}_{\mathrm{MS}}$ and has good power compared to all the other tests considered in the simulation study.

Model	C	Sample size (#ASPs)	Type I error (%)				
			Likelihood Ratio Tests		Wald Test	Score Tests	
			$\mathrm{LRT}_{\mathrm{MS}}$	$\mathrm{LRT}_{\mathrm{Li}}$		$T_{S\_\max}$, $T^B_{S\_\max}$	T_{S_sum}, $T^B_{S\_\mathrm{sum}}$
1-C	.12	200	4.9	3.3	1.0	4.8, 3.2	-
		1000	5.3	5.0	5.5	6.5, 4.7	-
1-D	.2	200	6.4	4.8	2.9	6.5, 4.9	-
		1000	5.8	5.3	8.1	8.2, 6.0	-
1-E	.3	200	8.0	5.7	6.7	8.1, 5.6	-
		1000	6.0	5.9	8.5	8.4, 6.3	-

Table 7.1: Type I error for tests of two-versus-one disease genes in a region for α = 5% (nominal). Type I errors of nominal 5% tests assuming asymptotic chi-squared distributions. The disease gene was located at 45 cM ($\tau = 45$). A marker map of 10 equally spaced markers covering a total of 90 cM was used (markers at 0,10,…,90cM). [Type I errors of T_{S_sum} and $T^B_{S\_\mathrm{sum}}$ are not estimated, because the asymptotic null distributions of these statistics are difficult to predict and do not approximate well defined distribution.]

Model	C	Sample size (#ASPs)	Estimated Critical Values for α = 5%				
			LRTs		Wald Test	Score Tests	
			$\mathrm{LRT}_{\mathrm{MS}}$	$\mathrm{LRT}_{\mathrm{Li}}$		$T_{S\_\max}$, $T^B_{S\_\max}$	T_{S_sum}, $T^B_{S\_\mathrm{sum}}$
1-C	.12	200	5.81	5.06	3.24	5.94, 5.38	8.37, 6.24
		1000	6.24	5.97	6.36	6.70, 5.88	8.31, 6.79
1-D	.2	200	7.12	5.86	4.69	6.62, 5.90	8.57, 6.50
		1000	6.43	6.06	7.54	7.36, 6.57	9.21, 7.15
1-E	.3	200	7.16	6.39	7.00	7.31, 6.42	9.44, 7.36
		1000	6.48	6.45	11.67	7.43, 6.39	9.67, 7.02

Table 7.2: Estimated 5% critical values for tests of two-versus-one disease genes in a region based on 1000 replicates.

Model	Penetrance matrix				Allele frequencies	Prevalence	τ_2- τ_1	C_1, C_2
2-A		aa	aA	AA	Pr(A) = 0.18	3.4%	20	0.10, 0.10
	bb	.002	.060	.060	Pr(B) = 0.18		40	0.08, 0.08
	bB	.060	.060	.060				
	BB	.060	.060	.060				
2-B		aa	aA	AA	Pr(A) = 0.064	3.4%	20	0.22, 0.22
	bb	.002	.140	.140	Pr(B) = 0.064		40	0.18, 0.18
	bB	.140	.140	.140				
	BB	.140	.140	.140				

Table 7.3: Two-locus models used for power study.

Model	τ_1, τ_2	C_1, C_2	Power (%)				
			LRTs		Wald Test	Score Tests	
			$\mathrm{LRT}_{\mathrm{MS}}$	$\mathrm{LRT}_{\mathrm{Li}}$		$T_{S\_\max}$, $T^B_{S\_\max}$	T_{S_sum}, $T^B_{S\_\mathrm{sum}}$
2-A	40,60	0.10, 0.10	24.8	25.3	15.0	26.3, 27.0	28.2, 26.5
	30,70	0.08, 0.08	61.3	61.6	30.6	55.3 , 57.4	60.2, 62.6
2-B	40,60	0.22, 0.22	98.2	97.9	81.9	98.6, 98.6	98.6, 98.7
	30,70	0.18, 0.18	83.1	100.0	99.5	100.0, 100.0	100.0, 100.0

Table 7.4: Approximate power of α = 5% tests of two-versus-one disease genes in a region. Critical values were approximated using results shown in Table 7.2. Sample size: 1000 ASPs. Marker map: 11 equally spaced markers spanning 0-100 cM.

7.3.2 Performance of the statistic $\mathrm{LRT}_{\mathrm{Li}}$

Observed type I errors of the statistic $\mathrm{LRT}_{\mathrm{Li}}$ are shown in Table 7.5 for several underlying one-locus models and sample sizes. Overall, the test appears to have type I errors close to the nominal 5%, although the type I errors may depend on the sample size, and the underlying one-locus disease model. Table 7.6 shows that the observed type I error of the test may also depend on the initial values selected for the GEE algorithm, once again demonstrating the importance of carefully choosing the initial values. The results suggest that placing initial location values (τ's) too far apart tends to increase the type I error.

Penetrance vector; Pr(D)	τ	C	Sample size (#ASPs)	Marker map (spacing)	Type I error (%)
					$\mathrm{LRT}_{\mathrm{Li}}$
(.01,.05,.05); 0.25	45	.105	1000	0-90 (10)	4.7
	25	.105	1000	0-90 (10)	5.8
(.01,.05,.05); 0.25	45	.105	200	0-90 (10)	5.7
			500	0-90 (10)	5.7
			1000	0-90 (10)	4.7
(.01,.03,.03); 0.3	45	.055	1000	0-90 (10)	5.3
(.01,.05,.05); 0.25	45	.105	1000	0-90 (10)	4.7
(.01,.1.,1); 0.2	45	.174	1000	0-90 (10)	4.9
(.01,.25,.25); 0.10	45	.299	1000	0-90 (10)	3.3

Table 7.5: Type I error for the $\mathrm{LRT}_{\mathrm{Li}}$ test of two-versus-one disease genes in a region. Using critical values from a chi-squared distribution with 2 df, for a nominal 5% α-level.

Penetrance vector; Pr(D)	τ	C	SS	Initial	Type I error (%)
			(#ASPs)	τ_1, τ_2	$\mathrm{LRT}_{\mathrm{Li}}$
(.01,.1.,1); 0.2	45	.174	1000	30,60	4.9
				35,55	3.3
(.01,.05,.05); 0.25	25	.105	1000	25,45	5.8
				15,35	4.0
(.01,.05,.05); 0.25	45	.105	500	30,60	7.1
				35,55	5.7
(.01,.05,.05); 0.25	45	.105	200	30,60	7.0
				35,55	5.7

Table 7.6: Type I error: Sensitivity to initial values.

Models used to study power of $\mathrm{LRT}_{\mathrm{Li}}$ are shown in Table 7.7. Table 7.8 reports power of $\mathrm{LRT}_{\mathrm{Li}}$ under several two-locus models, distances between the two disease genes, and sample sizes. By considering four different underlying genetic models, the first section of the table demonstrates the increase in power of the test as the expected IBD sharing at the two genes (C_1 and C_2) increases. The second section of the table shows the power loss that results when the two disease genes are in close proximity to one another. In the last three sub-sections of the table, the effect of sample size on power is demonstrated for three distances separating the two disease genes. The expected trend of increased power with increased sample size is evident. The results show that the impact of reduced sample size on reduction in power is more pronounced when the two disease genes are relatively close to one another (Figure 7.2).

Model	Penetrance matrix				Allele frequencies	τ_2- τ_1	C_1, C_2
2-C		aa	aA	AA	Pr(A) = 0.20	20	0.073, 0.073
	bb	.010	.100	.100	Pr(B) = 0.20		
	bB	.100	.100	.100			
	BB	.100	.100	.100			
2-D		aa	aA	AA	Pr(A) = 0.10	20	0.161, 0.161
	bb	.010	.250	.250	Pr(B) = 0.10		
	bB	.250	.250	.250			
	BB	.250	.250	.250			
2-E		aa	aA	AA	Pr(A) = 0.05	20	0.233, 0.233
	bb	.010	.400	.400	Pr(B) = 0.05		
	bB	.400	.400	.400			
	BB	.400	.400	.400			
2-F		aa	aA	AA	Pr(A) = 0.015	20	0.303, 0.303
	bb	.010	.600	.600	Pr(B) = 0.015		
	bB	.600	.600	.600			
	BB	.600	.600	.600			
2-G		aa	aA	AA	Pr(A) = 0.015	10	0.327, 0.327
	bb	.010	.010	.800	Pr(B) = 0.015	15	0.303, 0.303
	bB	.010	.010	.800		20	0.283, 0.283
	BB	.800	.800	.800		40	0.235, 0.235

Table 7.7: Two-locus models used to study power of $\mathrm{LRT}_{\mathrm{Li}}$.

Model	τ_1, τ_2	C_1, C_2	Sample size	Observed power of LRT$_{\text{Li}}$ (%)
2-C	40, 60	**0.073, 0.073**	1000	9.4
2-D	40, 60	**0.161, 0.161**	1000	84.2
2-E	40, 60	**0.233, 0.233**	1000	98.8
2-F	40, 60	**0.303, 0.303**	1000	99.3
2-G	**45, 55**	**0.327, 0.327**	1000	63.5
	42.5, 57.5	**0.303, 0.303**	1000	96.3
	40, 60	**0.283, 0.283**	1000	99.7
	30, 70	**0.235, 0.235**	1000	100
2-G	30, 70	0.235, 0.235	**200**	86.4
			500	99.9
			1000	100
	40, 60	0.283, 0.283	**200**	53.0
			500	94.3
			1000	99.7
	42.5, 57.5	0.303, 0.303	**200**	20.0
			500	69.8
			1000	96.3

Table 7.8: Power of LRT$_{\text{Li}}$. For each simulation the marker map consisted of 11 equally spaced markers on a map of 0-100 cM (i.e. markers at 0,10,…,100cM)

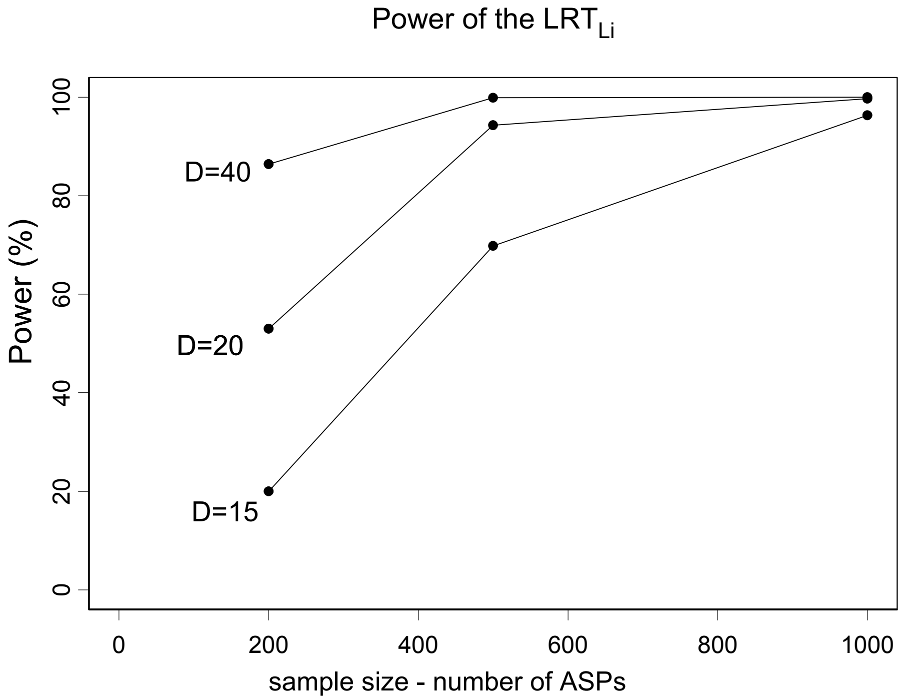


Figure 7.2: Power of LRT_{Li} (under model 2-G from Table 7.7). D = distance between two disease genes in cM.

7.4 Discussion

Although further comparisons of the test statistics are warranted, the limited simulations described in section 7.3.1 favour the quasi-likelihood ratio test LRT_{Li} because of its simplicity, more predictable distribution under the null hypothesis, and good power relative to the other tests under the alternative hypothesis. Further improvements to this test may be possible. In particular, Hanfelt and Liang (1995) showed that the projection approach of Li (1993) is closely related to an approach directly based on the quasi-likelihood and proposed extensions to improve the approximation in Li's method.

Recall that in our score tests, we computed the score statistics at the value of the nuisance location parameter estimated under the alternative hypothesis. Some alternative methods of constructing test statistics when a nuisance parameter cannot be estimated under the null hypothesis have been proposed. Davies suggested maximizing the statistic over the range of possible values for the parameter, and showed how to approximate an upper bound for the significance level (Davies 1977) or the significance probability (Davies 1987). Lemdani and Pons (1995) apply this method to the formulation of a likelihood ratio test statistic in a parametric linkage heterogeneity model. Alternatively, Liang and Rathouz (1999) propose a class of score statistics for the linkage heterogeneity model. They suggest finding the MLE of the nuisance parameter under the alternative hypothesis with a specific value of the parameter of interest, and computing the score statistic at this estimate of the nuisance parameter. Finally, Chiu et al. (2002) extended these linkage heterogeneity tests to apply to data from multiple markers, and discussed alternative ways of determining an appropriate null value of the location parameter. Some of these approaches may be applicable in designing tests for the number of disease genes in a region.

Further development of GEE parameter estimation in the two-locus model may lead to improved properties of the tests. For instance, properly constraining parameters can improve power. Note that in the δ_2-parameterized model, the null hypothesis from section 7.2.2

$$H_0\text{: } \tau_1 - \tau_2 = 0 \text{ or } C_1 = C_2 e^{-.04|\tau_2 - \tau_1|} \text{ or } C_2 = C_1 e^{-.04|\tau_2 - \tau_1|}.$$

could be written simply as

$$H_0\text{: } C_1 = C_2 e^{-.04|\tau_2 - \tau_1|} \text{ or } C_2 = C_1 e^{-.04|\tau_2 - \tau_1|}$$

had more stringent constraints been placed on the parameters, which would force $C_1 = C_2$ when $\tau_1 = \tau_2$.

Chapter 8

Application to Type 1 Diabetes Genome Scan Data

We applied the estimation and testing methods described in this thesis to previously analyzed data from a genome scan for type 1 diabetes (Mein et al. 1998) and obtained estimates of two putative disease gene locations on chromosome 6, approximately 20 cM apart, and two gene locations on chromosome 16, approximately 55 cM apart. A description of the data and results of the analyses are presented in this chapter.

8.1 Introduction

There is very strong evidence for at least one type 1 diabetes susceptibility gene (the IDDM1 locus) in the HLA/MHC region on chromosome 6, which explains about 34% of the familial clustering of the disease (Davies et al. 1994). Although (suggestive) evidence of linkage has been obtained in many other genomic regions, none of the other genes appear to have such large effects on disease susceptibility. Delépine et al. (1997) and Cordell et al. (2000) reported evidence of a second disease gene for type 1 diabetes linked to the HLA region in two different collections of families, although the positions they reported were about 20 cM apart. Using another set of ASPs, Concannon et al. (1998) also obtained suggestive evidence for linkage at the second gene location identified by Delépine et al. (1997).

The genome scan data analyzed here consists of genotypes for 356 ASP families with type 1 diabetes. In total, 355 microsatellite markers were genotyped. A detailed description of the data can be found in Davies et al. (1994), Cucca et al. (1998), and Mein et al. (1998). Analyses of chromosome 6 and 16 data will be described in sections 8.2.1 and 8.2.2, respectively. Chromosome 6 was chosen for analysis because of previous reports suggesting that two linked genes on chromosome 6 may be contributing to type 1 diabetes susceptibility (Delépine et al. 1997; Cordell et al. 2000; Concannon et al. 1998). Chromosome 16 was studied because the NPL curve contained two peaks, which could be the result of two distinct type 1 diabetes genes (Figure 8.3).

8.2 Analysis of Selected Chromosomes

8.2.1 Chromosome 6 analysis

The NPL plot for a 100cM segment of chromosome 6, based on genotype data at 18 markers (Figure 8.1), shows very strong evidence of linkage at approximately 25cM. This is the position of the well-established IDDM1 gene(s) in the MHC/HLA region (Davies et al.

1994). Figure 8.2 presents average IBD sharing at 18 chromosome 6 markers in the sample of 356 ASPs, along with expected IBD sharing curves defined by estimates from the one-locus and two-locus models. Assuming there is only one disease gene in the region, the location of the single putative disease gene was estimated to be 28.0cM, with a standard error of 0.74. Mean IBD sharing (C+1) was estimated to be 1.47 (s.e. 0.04). With the two-locus model, putative disease gene locations were estimated to be 27.0 cM (s.e. 0.52) and 48.6 cM (s.e. 5.64). Expected IBD sharing for an affected sib-pair was estimated to be 1.45 (s.e. 0.03) and 1.26 (s.e. 0.04) at these two loci, respectively. The one-locus-model asymptotic-theory 95% confidence interval for the disease gene location is (26.6,29.5). The 95% CIs for the disease gene locations under the two-locus model are (26.0,28.0) and (37.4,59.5). The confidence interval for the second gene location is wider because the effect size of the second putative gene is low relative to that of the first gene.

Test statistics for evaluating evidence for a single disease gene described in section 6.1 were calculated for the chromosome 6 data. The five statistics (LRT_{MS}, LRT_{Li}, Wald Test, Score Test, and L^*) were all highly significant with p-values < .0001. These p-values were obtained using asymptotic distributions rather than using empirical critical values. Simulations presented in chapter 6 indicated that all of these statistics have close to nominal significance levels. Furthermore, in this analysis, all five statistics were very large and would surely be significant following an empirical correction of critical values.

The approximate likelihood ratio test statistic LRT_{Li} was calculated to assess evidence for a second disease gene in the region. Using the asymptotic chi-squared distribution with 2 degrees of freedom, the test was significant at the 5% level ($LRT_{Li} = 8.86$, $p_{asymp} = 0.01$). By generating 1000 fully informative data sets of 356 ASPs under a null one-locus model resembling the observed one-locus model, the empirical p-value of the likelihood ratio statistic was approximated to be 0.08. Note that, first of all, this is not a precise approximation since it was based on only 1000 data sets (95% CI for p-value: 0.063 to 0.097). Also this p-value is a conservative estimate of significance, because the simulated data was fully informative yielding higher test statistics which would lead to a higher critical value. As a result, empirical p-values for the data will tend to be too large. Thus, it can be concluded that there is suggestive evidence for a second type 1 diabetes susceptibility gene, linked to the HLA region, approximately 20-30 cM centromeric from IDDM1.

8.2.2 Chromosome 16 analysis

The chromosome 16 data consisted of genotypes in a sample of 351 ASPs for 20 markers spanning approximately 110 cM. The shape of the NPL plot for chromosome 16 suggests that there may be two disease genes on this chromosome that influence type 1 diabetes susceptibility (Figure 8.3). Assuming there is only one disease gene in the region, and applying the GEE method described in section 2.2, the location of the single putative disease gene was estimated to be 89.4, with an estimated standard error of 2.9. Expected IBD sharing (C+1) was estimated to be 1.14 (s.e.= 0.04). When the two-locus estimation method proposed in chapter 4 was applied to the same data, the two putative disease gene locations were estimated to be 37.0 cM (s.e. 9.9) and 91.6 cM (s.e. 3.6). Expected IBD sharing in affected sib-pairs was estimated to be 1.07 (s.e.= 0.04) and 1.14 (s.e.= 0.04) at these two loci, respectively. Figure 8.4 shows the average marker IBD sharing of the 351 ASPs from the sample along with curves defined by estimates from the one-locus and two-locus models.

The one-locus model asymptotic-theory 95% confidence interval for the disease gene location is (83.8,95.0). The 95% CIs for the two disease gene locations under the two-locus model are (17.6,56.5) and (84.5,98.6). The CI for the first disease gene location is quite wide, because the excess sharing in this part of the chromosome is actually not very high, and there is little evidence for linkage to a second disease gene in this region of the chromosome. Simulation results presented in chapter 5 showed that gene location confidence intervals can have lower than nominal coverage, particularly for genes with small effect sizes. Therefore, although the two gene location confidence intervals for chromosome 16 do not overlap, this cannot be taken as evidence for two distinct type 1 diabetes susceptibility genes on this chromosome.

We applied tests to assess evidence for one and two type 1 diabetes susceptibility genes on chromosome 16. There was significant evidence for one disease gene in the region (Table 8.1). The test statistic for two versus one disease genes in the region based on the method of Li (1993) was not significant (LRT_{Li} = 0.554, p_{asymp} = 0.76). Power of the likelihood ratio test LRT_{Li} to detect the second gene in the estimated two-locus model was

Test	Test Statistic	p-value[a]
LRT_{MS}	13.86	0.001
LRT_{Li}	12.10	0.002
Wald Test	13.38	0.001
Score Test	14.38	0.001
L^*	2.606	0.005

Table 8.1: Tests for one-versus-zero disease genes on chromosome 16

[a] p-values based on asymptotic distributions

approximated for several sample sizes (Figure 8.5). It is clear that even if there is a second disease gene with mean IBD sharing of about 0.07 located at 37cM on this chromosome, as suggested by the two-locus-model estimates, it would not be detectable by the proposed methods. In fact, it would take a sample of about 1200 ASPs to reach 80% power to detect this gene. A gene with such low mean IBD sharing would have a low probability of detection by model-free linkage analysis methods with typical sample sizes of less than 1000 ASPs.

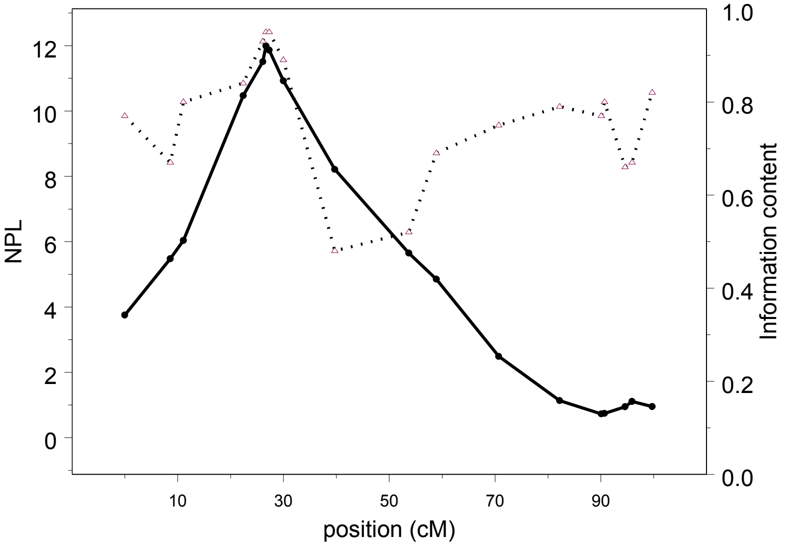


Figure 8.1: NPL plot for a 100cM region of chromosome 6 based on type 1 diabetes genome scan data (Mein et al. 1998). The solid line represents the NPL curve. The dashed line shows the marker information content (scale on right side of plot).

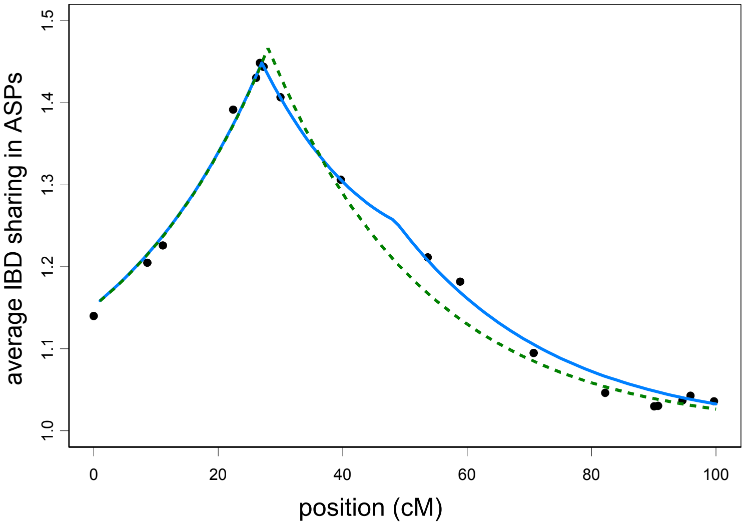


Figure 8.2: Average IBD sharing at 18 chromosome 6 markers in a sample of ASPs with diabetes (data: Mein et al. 1998). The two curves show the expected IBD sharing for the estimated one-locus model (dashed line) with C=0.467 and τ=28.0, and the two-locus model (solid line) with C_1=0.449, C_2=0.256, τ_1=27.0, and τ_2=48.5.

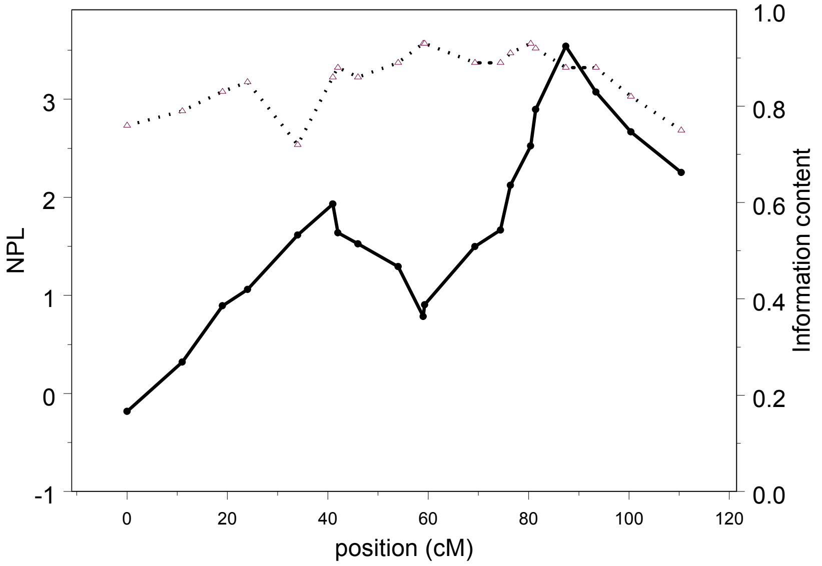


Figure 8.3: NPL plot for chromosome 16 based on type 1 diabetes genome scan data (Mein et al. 1998). The solid line represents the NPL curve. The dashed line shows the marker information content (scale on right side of plot).

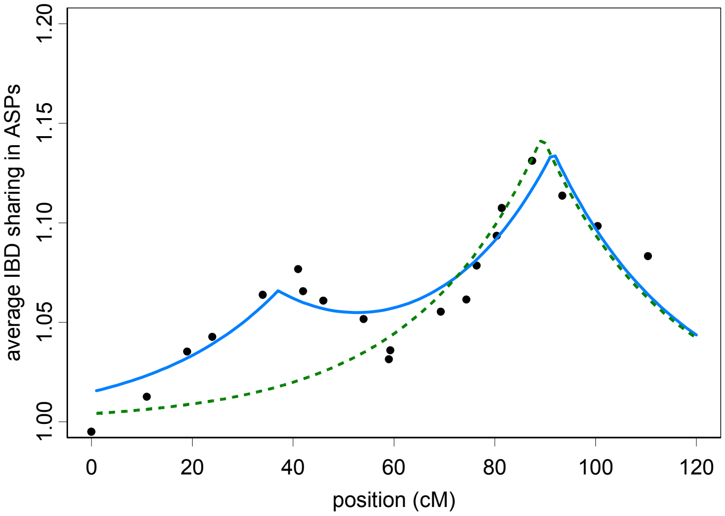


Figure 8.4: Average IBD sharing at 20 chromosome 16 markers in a sample of ASPs with diabetes (data: Mein et al. 1998). The two curves show the expected IBD sharing for the estimated one-locus model (dashed line) with C=0.143 and τ=89.4, and the two-locus model (solid line) with C_1=0.066, C_2=0.136, τ_1=37.0, and τ_2=91.6.

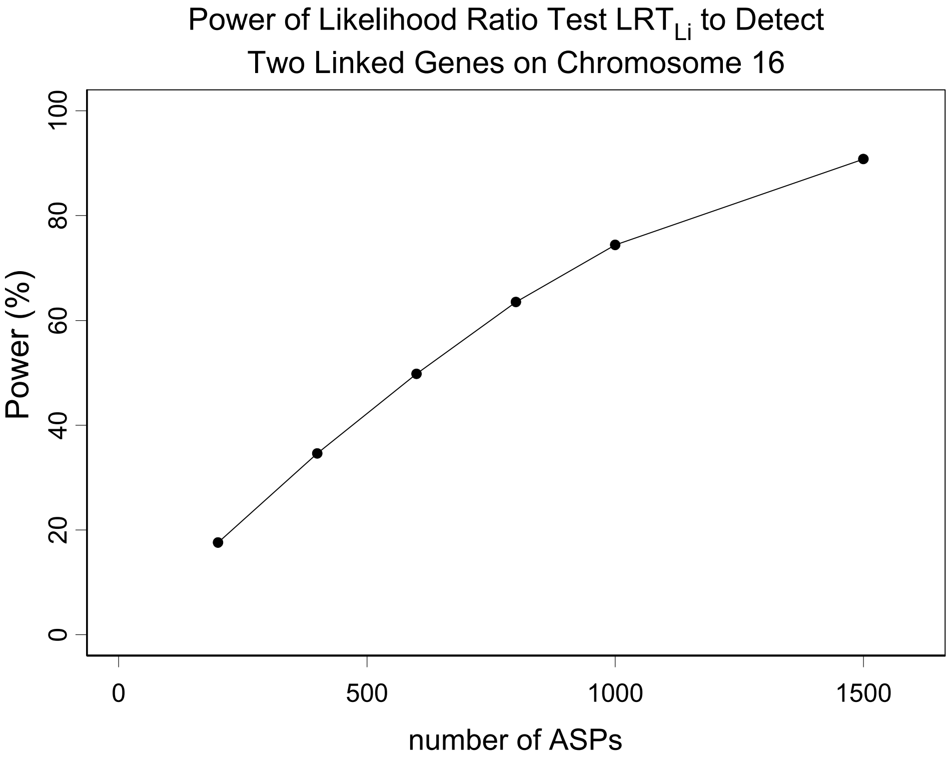


Figure 8.5: Power curve for detecting two putative disease genes on chromosome 16.

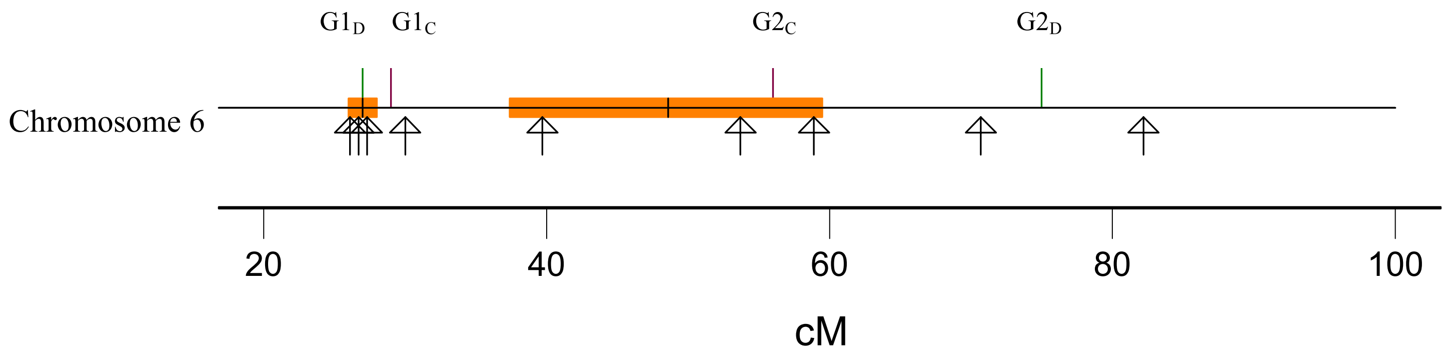


Figure 8.6: Comparison of our chromosome 6 results to those from previous studies. Arrows show marker positions in our analysis. G1$_D$ and G2$_D$ are the two gene locations identified by the analysis of Delépine et al. (1997), and G1$_C$ and G2$_C$ are the two genes identified by the analysis of Cordell et al. (2000). Also shown are the confidence intervals (filled boxes) and the point estimates (vertical lines within the filled boxes) for the two gene locations from our analysis.

8.3 Discussion of Results

Cordell et al. (2000) obtained evidence for a second type 1 diabetes gene on chromosome 6 approximately 30cM from IDDM1. Delépine et al. (1997) also concluded that there is a second type 1 diabetes susceptibility gene on chromosome 6 linked to HLA, although they suggested it is located almost 50 cM from IDDM1. A diagram showing the results of our analysis compared to the results of Cordell et al. and Delépine et al. is shown in Figure 8.6. Our confidence interval for the second disease gene location includes the location for the second gene reported by Cordell et al. (2000), but not the second gene reported by Delépine et al. (1997). However, in light of the simulation results presented in chapter 5, we conclude that the confidence interval for the second gene may be too narrow, and hence to have less than nominal 95% coverage. It should also be noted that the data we analyzed is the same data set that was analyzed by Cordell et al. (2000). Another difference between our analysis and that of Delépine et al. is that our method does not account for sex-specific recombination fractions. It has been shown that recombination fractions in the chromosome 6 region we studied are different between male and female meioses. The analysis carried out by Delépine et al. (1997) did take this difference into account, which may partially explain the difference in location estimates of the second gene.

When there are two disease genes in one chromosomal region contributing to the susceptibility of the same disease, our GEE estimation method can improve their localization. In the analysis presented in section 8.2.1, the standard error of the estimated location of the gene in the HLA region under the two-locus model is smaller than the standard error for the location under the one-locus model. Furthermore, in comparison to the estimated location under the one-locus model, the estimated location of IDDM1 under the two-locus model is closer to the D6S273 and TNFa marker loci commonly sited as the IDDM1 location. Thus, apparently, a more accurate and more precise estimate of the major IDDM1 gene was obtained by fitting the two-locus model. This improvement in localization is confirmed by our simulation studies.

In this type of situation, with one of the genes having a substantially larger effect than the other, conditional methods such as that of Farrall (1997) may work well. In such cases, the first (major) gene may be successfully localized with minimal bias associated with ignoring effects of the second (minor) gene. However, if the effects of two genes on IBD sharing are not drastically different, we believe there would be greater benefits of using our method, which simultaneously estimates both disease gene locations.

Analysis of chromosome 16 data did not provide significant evidence for two linked disease genes. Note, however, that if there are two genes at the locations identified by our analysis (approximately 37 and 92 cM), then these two genes are quite far apart. Because the linkage between these two genes is low, methods that do not account for the existence of multiple linked disease genes may perform reasonably well in localizing the two putative genes. However, the performance of the two-locus model is not compromised by lack of linkage between the two disease genes.

In summary, we have applied the estimation and testing methods for the detection of two linked disease genes to chromosome 6 and chromosome 16 data from a type 1 diabetes genome scan. The analyses demonstrated that when two linked disease genes are (apparently) present in one region, they can be localized with greater accuracy and precision by applying the two-locus-model GEE method. The methods described in this thesis would presumably be most useful for localizing disease genes separated by moderate genetic distances when neither of the genes has a drastically larger effect on IBD sharing than the other.

Chapter 9

Conclusions

The main scientific contribution of this thesis, development of a method for simultaneous localization of two linked disease genes, was presented in chapter 4. In addition, several test statistics that can be used in conjunction with this estimation method were proposed. Properties of the estimation and testing methods were studied by simulation, and the methods were applied to a type 1 diabetes data set. As background to the development of the model for two linked disease genes, the generalized estimating equations approach for localizing a single disease gene in a chromosomal region proposed by Liang et al. (2001*a*) was examined in chapter 3.

Several modifications of the one-locus estimating equations were considered in chapter 3. Simulations demonstrated that while more careful modeling of the covariance structure of the data can improve estimation efficiency, unbiased location estimates can be obtained by assuming a simple working correlation and empirically estimating the variances of IBD sharing at each marker. Based on these observations, there is reason to believe that using a similar approach may perform well in the estimation of two disease genes in one region, where the true covariance structure of the data is more complicated than in the one-locus model. In section 3.3 the GEE method for localizing a single disease gene was extended to allow the estimation of IBD sharing probabilities for ASPs at a single disease gene. This extension not only has the potential to improve disease gene localization under certain underlying genetic models, but it also provides estimates of parameters which are useful in evaluating the significance of tests for the number of disease genes in a region. Some of the test statistics introduced in chapter 7 do not have well defined distributions under the null hypothesis. Empirical critical values must then be approximated by simulation. This is particularly problematic in the case of testing for two versus one disease genes in a region. Estimates of IBD sharing probabilities, obtained, for example, by the method described in section 3.3, can be used in this case to identify an appropriate one-locus model for simulating data under the null hypothesis.

In chapter 4, the GEE approach for localizing a single disease gene proposed by Liang et al. (2001*a*) and modified in chapter 3, was extended to a method for localizing two linked disease genes. We consider whether performance of the estimation algorithm could perhaps be improved by correctly modeling the covariance matrix of the data for each ASP under the two-locus model. Simulations showed, however, that even with the simple covariance applied in an original implementation, with a sufficiently informative sample, the proposed GEE method provides unbiased estimates with confidence interval coverage near the nominal level. The method generally produces better results when the two disease loci are further apart, when there is higher allele sharing at the two disease genes, and when more ASPs are available for analysis. When the two disease genes are believed to be located close to one another, a denser map of markers is recommended, to ensure that there are several markers (3-4) between the two disease genes. As the estimation algorithm can be quite sensitive to initial parameter values under certain models, we recommend carefully choosing initial values, for example, based on NPL or mean allele-sharing plots. Also, a sensitivity analysis can be performed to study the impact of initial values on parameter estimates for a given data set.

One of the benefits of this method is that confidence intervals can be computed using the location estimates and their robust standard errors. Although for some underlying genetic disease models coverage of asymptotic theory confidence intervals reached the nominal levels with a sufficiently large sample, for other simulation settings coverage was below the nominal level. In light of these findings, we caution that in the diabetes data application presented in this thesis (and in similar practical situations), the confidence interval for the second putative disease gene on chromosomes 6 and 16 may not have the nominal 95% coverage. In this study, the actual coverage may be lower because neither the sample size (356 ASPs) nor the estimated effect sizes of the secondary genes were large.

Farrall (1997) described a likelihood approach for testing the independent support of a putative susceptibility gene that maps close to a previously established gene. In his approach it was assumed that an “anchor” gene has already been mapped, and evidence for linkage at linked loci was evaluated taking into account the first mapped gene. This conditional evidence for linkage was then used as the basis for mapping a second disease gene linked to the “anchor” gene. However, as our results and the results of other studies (Hauser et al. 2003) have shown, localization of the first gene may be incorrect if the presence of other linked disease genes is not taken into account. “Conditional approaches” for mapping a second disease gene linked to an already mapped disease gene may be suitable in situations when the first gene has a much stronger effect than the second linked gene, as is the case for chromosome 6 in the diabetes data analyzed in chapter 8. However, if both genes have similar effect size (or the difference in effect size is not large), these approaches would probably be inferior to an approach that maps the two genes simultaneously. In such cases, our methods may localize the two genes more precisely. Further research comparing conditional and simultaneous gene searches is needed.

Under some basic assumptions the GEE method for localizing genes can be easily extended to allow estimation of three or more linked disease genes. We present the mean sharing function and briefly describe this extension in Appendix D. However, there are likely to be more computational problems as the number of linked disease genes is increased.

Several tests for assessing the evidence for one or two disease genes in a region were developed and evaluated by simulation. The simplicity of the approximated quasi-likelihood ratio tests and simulation results favour the approximate quasi-likelihood ratio test statistic based on the method of Li (1993). Further research aimed at comparing all proposed statistics and identifying more powerful tests under a variety of genetic models and study designs would be of great value. More extensive simulations are necessary to assess performance of the tests presented in this thesis and perhaps new test statistics for a broader range of underlying genetic models, sample sizes, and marker maps.

Several suggestions have been made in this and preceding chapters, which may improve the estimation and testing methods introduced in this thesis. Placing constraints on the parameters may lead to further improvements, particularly in terms of increasing power of the tests. Holmans (1993) introduced a set of constraints for the IBD sharing probabilities of affected sib pairs under a single disease gene model. In the single locus model parameterized in terms of mean IBD sharing (C) these constraints translate to $0 \leq C \leq 1$. Similar constraints could be imposed on the two expected excess IBD sharing parameters (C_1,C_2), taking into account the fact that there are two disease genes separated by a given genetic distance. Another area where improvements could be made is in the mean sharing model itself, which could be formulated as a function of male and female recombination fractions, rather than a combined estimate. Although it is not uncommon to assume equal recombination fractions between sexes in linkage analysis, there is evidence that this assumption is violated, particularly in some specific regions of the genome, including the HLA region which contains the type 1 diabetes gene IDDM1. Thus, accounting for inequality of male and female recombination fractions has the potential to improve the precision of disease gene location estimates. Finally, extension of the GEE gene localization method to account for the uncertainty and bias of estimated IBD sharing, S^*, when markers are non-fully informative, would be useful.

In conclusion, we have introduced, applied, and evaluated methods for estimating the locations of two linked disease susceptibility genes. In addition we have suggested test statistics that may assist in determining the number of disease genes (0, 1, or 2) in a region. The estimation method can provide unbiased estimates of disease gene locations, while avoiding multiple testing problems by looking at all markers jointly. Because this method provides confidence intervals for the two gene locations, it can help define region(s) for further studies aimed at fine-mapping of the trait loci. Furthermore, the proposed test statistics can help prioritize regions for further study.

References

1. Aitchison J, Silvey SD (1960) Maximum-likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society. Series B (Methodological)* 22: 154-171.
2. Biswas S, Papachristou C, Irwin M, Lin S (2003) Linkage analysis of the simulated data – evaluations and comparisons of methods. *BMC Genetics* 4(Suppl 1): S70.
3. Boos DD (1992) On generalized score tests. *The American Statistician* 46: 327-333.
4. Breslow N (1990) Tests of hypotheses in overdispersed poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* 85: 565-571.
5. Casella G, Berger RL (1990) Statistical Inference. Duxbury Press. Belmont, California.
6. Chiu Y-F, Liang K-Y (2004) Conditional multipoint linkage analysis using affected sib pairs: An alternative approach. *Genetic Epidemiology* 26: 108-115.
7. Chiu Y-F, Liang K-Y, Beaty TH (2002) Multipoint linkage detection in the presence of heterogeneity. *Biostatistics* 3: 195-211.
8. Concannon P, Gogolin-Ewens KJ, Hinds DA, Wapelhorst B, Morrison VA, Stirling B, Mitra M, Farmer J, Williams SR, Cox NJ, Bell GI, Risch N, Spielman RS (1998) A

second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nature Genetics* 19(3): 292-296.

9. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11: 2463-2468.
10. Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M (1995) Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type I diabetes. *The American Journal of Human Genetics* 57: 920-934.
11. Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, Wicker LS, Clayton DG (2001) Statistical modeling of interlocus interactions in a complex disease: Rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* 158: 357-367.
12. Cordell HJ, Todd JA, Lathrop GM (1998) Mapping multiple linked quantitative trait loci in non-obese diabetic mice using a stepwise regression strategy. *Genetical Research* 71: 51-64.
13. Cordell HJ, Wedig GC, Jacobs KB, Elston RC (2000) Multilocus linkage tests based on affected relative pairs. *The American Journal of Human Genetics* 66: 1273-1286.
14. Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A (1999) Loci on chromosome 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics* 21: 213-215.
15. Cucca F, Esposito L, Goy JV, Merriman ME, Wilson AJ, Reed PW, Bain SC, Todd JA (1998) Investigation of linkage of chromosome 8 to type 1 diabetes: multipoint analysis and exclusion mapping of human chromosome 8 in 593 affected sib-pair families from the U.K. and U.S.. *Diabetes* 47: 1525-1527.

16. Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farrall M, Barnett AH, Bain SC, Todd JA (1994) A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* 371:130-6.

17. Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64: 247-254.

18. Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74: 33-43.

19. Delépine M, Pociot F, Habita C, Hashimoto L, Froguel P, Rotter J, Cambon-Thomsen A, Deschamps I, Djoulah S, Weissenbach J, Nerup J, Lathrop M, Julier C (1997) Evidence of a non-MHC susceptibility locus for type 1 diabetes linked to HLA on chromosome 6. *The American Journal of Human Genetics* 60: 174-187.

20. Drum M and McCullagh P (1993) [Regression models for discrete longitudinal responses]: Comment. *Statistical Science* 8: 300-301.

21. Elston RC (2000) Introduction and overview. Statistical Methods in Genetic Epidemiology. *Statistical Methods in Medical Research* 9: 527-541.

22. Farrall M (1997) Affected sibpair linkage tests for multiple linked susceptibility genes. *Genetic Epidemiology* 14: 103-115.

23. Ghosh S, Watanabe RM, Hauser ER, Valle T, Magnuson VL, Erdos MR, Langefeld CD, Balow J Jr, Ally DS, Kohtamaki K, Chines P, Birznieks G, Kaleta HS, Musick A, Te C, Tannenbaum J, Eldridge W, Shapiro S, Martin C, Witt A, So A, Chang J, Shurtleff B, Porter R, Kudelko K, Unni A, Segal L, Sharaf R, Blaschak-Harvan J, Eriksson J, Tenkula T, Vidgren G, Ehnholm C, Tuomilehto-Wolf E, Hagopian W, Buchanan TA, Tuomilehto

J, Bergman RN, Collins FS, Boehnke M. (1999) Type 2 diabetes: Evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proceedings of the National Academy of Sciences* 96: 2198-2203.

24. Glidden DV, Liang KY, Chiu YF, Pulver AE (2003) Multipoint affected sibpair linkage methods for localizing susceptibility genes of complex diseases. *Genetic Epidemiology* 24: 107-117.
25. Goldin LR and Weeks DE (1993) Two-locus models of disease: Comparison of likelihood and nonparametric linkage methods. *The American Journal of Human Genetics* 53: 908-915.
26. Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM (1993) An Introduction to Genetic Analysis. New York: W. H. Freeman and Company.
27. Gudbjartsson DF, Jonasson K, Frigge M, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* 25: 12-13.
28. Hampe J, Frenzel H, Mirza MM, Croucher PJ, Cuthbert A, Mascheretti S, Huse K, Platzer M, Bridger S, Meyer B, Nurnberg P, Stokkers P, Krawczak M, Mathew CG, Curran M, Schreiber S (2002) Evidence for a NOD2-independent susceptibility locus for inflammatory bowel disease on chromosome 16p. *Proceedings of the National Academy of Sciences of the United States of America* 99: 321-326.
29. Hanfelt JJ, Liang K-Y (1995) Approximate likelihood ratios for general estimating functions. *Biometrika* 82: 461-477.
30. Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* 2: 3-19.

31. Hauser ER, Bass M, Martin ER (2003) Identification of gene locations from maximum likelihood ASP linkage analysis: Are there features of the lod score curve that distinguish regions with two loci? *The American Journal of Human Genetics* 73(5)supplement: 615.
32. Hodge SE, Vieland VJ (1996) The essence of single ascertainment. *Genetics* 144: 1215-1223.
33. Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. *The American Journal of Human Genetics* 52: 362-374.
34. Holmans P (2002) Detecting gene-gene interactions using affected sib pair analysis with covariates. *Human Heredity* 53: 92-102.
35. Hössjer O (2003) Assessing accuracy in linkage analysis by means of confidence regions. *Genetic Epidemiology* 25:59-72.
36. Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Le Cam L, Neyman J (ed). Proceedings of the Fifth Berkley Symposium in Mathematical Statistics and Probability. Volume 1. Berkeley: University of California Press. p 221-233.
37. Knapp M, Seuchter SA, Baur MP (1994) Two-locus disease models with two marker loci: The power of affected sib pair tests. *The American Journal of Human Genetics* 55: 1030-1041.
38. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: A unified multipoint approach. *The American Journal of Human Genetics* 58: 1347-1363.
39. Kruglyak L, Lander ES (1995) High-resolution genetic mapping of complex traits. *The American Journal of Human Genetics* 56: 1212-23.

40. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* 84: 2363-2367.

41. Lemdani M, Pons O (1995) Tests for genetic linkage and homogeneity. *Biometrics* 51: 1033-1041.

42. Li B (1993) A deviance function for the quasi-likelihood method. *Biometrika* 80: 741-753.

43. Liang K-Y, Chiu YF, Beaty TH (2000) A unified sampling approach for multipoint analysis of qualitative and quantitative traits in sib pairs. *The American Journal of Human Genetics* 66: 1631-1641.

44. Liang K-Y, Chiu YF, Beaty TH (2001*a*) A robust identity-by-descent procedure using affected sib-pairs: multipoint mapping for complex diseases. *Human Heredity* 51: 64-78.

45. Liang K-Y, Chiu YF, Beaty TH, Wjst M (2001*b*) Multipoint analysis using affected sib pairs: Incorporating linkage evidence from unlinked regions. *Genetic Epidemiology 21*: 105-122.

46. Liang K-Y, Rathouz PJ (1999) Hypothesis testing under mixture models: Application to genetic linkage analysis. *Biometrics* 55: 65-74.

47. Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13-22.

48. Lin S (2000) Monte Carlo methods for linkage analysis of two-locus disease models. *Annals of Human Genetics* 64: 519.532.

49. Lin S (2002) Construction of a confidence set of markers for the location of a disease gene using affected sib pair data. *Human Heredity* 53: 103-112.

50. Lipsitz SR, Fitzmaurice GM, Orav EJ, Laird NM (1994) Performance of generalized estimating equations in practical situations. *Biometrics* 50: 270.278.

51. Luo D-F, Buzzetti R, Rotter JI, Maclaren NK, Raffel LJ, Nistico L, Giovannini C, Pozzilli P, Thomson G, She J-X (1996) Confirmation of three susceptibility genes to insulin-dependent diabetes mellitus: IDDM4, IDDM5 and IDDM8. *Human Molecular Genetics* 5: 693-698.

52. MacLean CJ, Sham PC, Kendler KS (1993) Joint linkage of multiple loci for a complex disease. *The American Journal of Human Genetics* 53: 353-366.

53. Mancl LA, DeRouen TA (2001) A covariance estimator for GEE with improved small-sample properties. *Biometrics* 57: 126-134.

54. McCullagh P (1983) Quasi-likelihood functions. *Annals of Statistics* 11: 59-67.

55. McCullagh P (1991) Quasi-likelihood and estimating functions. In Hinkley DV, Reid N, Snell EJ (ed). Statistical Theory and Modelling. In Honour of Sir David Cox, FRS. London: Chapman and Hall.

56. McCullagh P and Nelder JA (1989) Generalized Linear Models. London: Chapman and Hall.

57. McLeish DL , Small CG (1992) A projected likelihood function for semiparametric models. *Biometrika* 79: 93-102.

58. Mein CA, Esposito L, Dunn MG, Johnson GC, Timms AE, Goy JV, Smith AN, Sebag-Montefiore L, Merriman ME, Wilson AJ, Pritchard LE, Cucca F, Barnett AH, Bain SC, Todd JA (1998) A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nature Genetics* 19(3) 297-300.

59. O'Hara Hines RJ (1998) Comparison of two covariance structures in the analysis of clustered polytomous data using generalized estimating equations. *Biometrics* 54: 312-316.

60. Olson JM, Witte JS, Elston RC (1999) Tutorial in biostatistics: Genetic mapping of complex traits. *Statistics in Medicine* 18: 2961-2981.

61. Ott J (1996) Complex traits on the map. *Nature* 379: 772-773.

62. Pan W (2001) On the robust variance estimator in generalised estimating equations. *Biometrika* 88: 901-906.

63. Pan W, Wall MM (2002) Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* 21: 1429-1441.

64. Prentice RL (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44: 1033-1048.

65. Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus Models. *The American Journal of Human Genetics* 46: 222-228.

66. Rotnizky A, Jewell NP (1990) Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77: 485-497.

67. Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: A powerful strategy for mapping complex genetic traits. *The American Journal of Human Genetics* 53: 1127-1136.

68. Sham P (1998) Statistics in Human Genetics. New York: Wiley.

69. Sham PC, MacLean CJ, Kendler KS (1994) Two-locus versus one-locus lods for complex traits. *The American Journal of Human Genetics* 55: 855-856.

70. Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP (2000) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: Application to mite sensitization. *The American Journal of Human Genetics* 66: 1945-1957.

71. Thomson EA (1975) The estimation of pairwise relationships. *Annals of Human Genetics* 39: 173-188.

72. Tiwari HK, Elston RC (1997) Linkage of multilocus components of variance to polymorphic markers. *Annals of Human Genetics* 61: 253-261.

73. Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61: 439-447.

74. White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1-25.

75. Wilson AF, Bailey-Wilson JE, Pugh EW, Sorant AJM (1996) The G.A.S.P.: a software tool for testing and investigating methods in statistical genetics. *The American Journal of Human Genetics* Suppl 59: A193.

76. Zeger SL, Liang K-Y (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42: 121-130.

77. Zinn-Justin A, Abel L (1998) Two-locus developments of the weighted pairwise correlation method for linkage analysis. *Genetic Epidemiology* 15: 491-510.

Appendix

Appendix A: Details of derivation of E(*S*(*t*)|Φ) under the two-locus model

Recall that

$$E[S(t)\mid\Phi]=\Pr(S(t)=1\mid\Phi)+2\Pr(S(t)=2\mid\Phi).$$

In chapter 4 (pg. 70-71) it was shown that, for $t<\tau_1<\tau_2$ (Case I)

$$\Pr(S(t)=j\mid\Phi)=\sum_{k=0}^{2}\Pr(S(t)=j\mid S(\tau_1)=k)\Pr(S(\tau_1)=k\mid\Phi),$$

and we stated that substituting the expressions for conditional IBD sharing probabilities from Table 3.7 and simplification leads to:

$$E[S(t)\mid\Phi]=1+(2\Psi_1-1)C_1 \text{ where } C_1\text{=E[}S(\tau_1)\text{|}\Phi\text{]-1 and } \Psi_1=\Psi_{t,\tau_1}=\theta_{t,\tau_1}{}^2+(1-\theta_{t,\tau_1})^2.$$

Some additional steps in the simplification are shown below:

$$\begin{aligned}E[S(t)\mid\Phi]&=\Pr(S(t)=1\mid\Phi)+2\Pr(S(t)=2\mid\Phi)\\
&=\Pr(S(t)=1\mid S(\tau_1)=0)\Pr(S(\tau_1)=0\mid\Phi)\\
&\quad+\Pr(S(t)=1\mid S(\tau_1)=1)\Pr(S(\tau_1)=1\mid\Phi)\\
&\quad+\Pr(S(t)=1\mid S(\tau_1)=2)\Pr(S(\tau_1)=2\mid\Phi)\\
&\quad+2\Pr(S(t)=2\mid S(\tau_1)=0)\Pr(S(\tau_1)=0\mid\Phi)\\
&\quad+2\Pr(S(t)=2\mid S(\tau_1)=1)\Pr(S(\tau_1)=1\mid\Phi)\\
&\quad+2\Pr(S(t)=2\mid S(\tau_1)=2)\Pr(S(\tau_1)=2\mid\Phi)\end{aligned}$$

$$
\begin{aligned}
&=\left[2\Psi_1(1-\Psi_1)+2(1-\Psi_1)^2\right]\left\{1-\Pr\left(S(\tau_1)=1\,|\,\Phi\right)-\Pr\left(S(\tau_1)=2\,|\,\Phi\right)\right\}\\
&\quad+\left[1-2\Psi_1(1-\Psi_1)+2\Psi_1(1-\Psi_1)\right]\Pr\left(S(\tau_1)=1\,|\,\Phi\right)\\
&\quad+\left[2\Psi_1(1-\Psi_1)+2\Psi_1{}^2\right]\Pr\left(S(\tau_1)=2\,|\,\Phi\right)\\
&=2-2\Psi_1+(2\Psi_1-1)\Pr\left(S(\tau_1)=1\,|\,\Phi\right)+(4\Psi_1-2)\Pr\left(S(\tau_1)=2\,|\,\Phi\right)\\
&=1-(2\Psi_1-1)+(2\Psi_1-1)\left\{\Pr\left(S(\tau_1)=1\,|\,\Phi\right)+2\Pr\left(S(\tau_1)=2\,|\,\Phi\right)\right\}\\
&=1+(2\Psi_1-1)\left\{E\left[S(\tau_1)\,|\,\Phi\right]-1\right\}\\
&=1+\left(2\Psi_1-1\right)C_1
\end{aligned}
$$

A similar derivation applies to $\tau_1 < \tau_2 < t$ (Case II).

For $\tau_1 < t < \tau_2$ (Case III):

$$
\begin{aligned}
E\left[S(t)\,|\,\Phi\right]&=\Pr\left(S(t)=1\,|\,\Phi\right)+2\Pr\left(S(t)=2\,|\,\Phi\right)\\
&=\sum_{k=0}^{2}\sum_{l=0}^{2}\frac{\Pr\left(S(t)=1\,|\,S(\tau_1)=k\right)\Pr\left(S(\tau_2)=l\,|\,S(t)=1\right)g_{kl}}{\Pr\left(S(\tau_2)=l\,|\,S(\tau_1)=k\right)}\\
&\quad+2\sum_{k=0}^{2}\sum_{l=0}^{2}\frac{\Pr\left(S(t)=2\,|\,S(\tau_1)=k\right)\Pr\left(S(\tau_2)=l\,|\,S(t)=2\right)g_{kl}}{\Pr\left(S(\tau_2)=l\,|\,S(\tau_1)=k\right)}
\end{aligned}
$$

where $g_{kl}=\Pr\left(S(\tau_1)=k,S(\tau_2)=l\,|\,\Phi\right)$. Let $\Psi_1=\Psi_{t,\tau_1}$, $\Psi_2=\Psi_{t,\tau_2}$, and $\Psi_3=\Psi_{\tau_1,\tau_2}$ where $\Psi=\theta^2+(1-\theta)^2$, θ being the recombination fraction between two loci. Substituting in conditional sharing probabilities from Table 3.7, and collecting terms for each g_{kl}, k = 0,1,2, l = 0,1,2 leads to:

$$\begin{aligned}
&E\left[S(t)\,|\,\Phi\right]\\
&= g_{00}\left[2\Psi_1(1-\Psi_1)\Psi_2(1-\Psi_2)+2(1-\Psi_1)^2(1-\Psi_2)^2\right]\Big/\Psi_3{}^2\\
&+g_{01}\left[2\Psi_1(1-\Psi_1)[1-2\Psi_2(1-\Psi_2)]+2(1-\Psi_1)^2 2\Psi_2(1-\Psi_2)\right]\Big/2\Psi_3(1-\Psi_3)\\
&+g_{02}\left[2\Psi_1(1-\Psi_1)\Psi_2(1-\Psi_2)+2(1-\Psi_1)^2\Psi_2{}^2\right]\Big/(1-\Psi_3)^2\\
&+g_{10}\left[[1-2\Psi_1(1-\Psi_1)]\Psi_2(1-\Psi_2)+\Psi_1(1-\Psi_1)(1-\Psi_2)^2\right]\Big/\Psi_3(1-\Psi_3)\\
&+g_{11}\left[[1-2\Psi_1(1-\Psi_1)][1-2\Psi_2(1-\Psi_2)]+\Psi_1(1-\Psi_1)2\Psi_2(1-\Psi_2)\right]\Big/[1-2\Psi_3(1-\Psi_3)]\\
&+g_{12}\left[[1-2\Psi_1(1-\Psi_1)]\Psi_2(1-\Psi_2)+\Psi_1(1-\Psi_1)\Psi_2{}^2\right]\Big/\Psi_3(1-\Psi_3)\\
&+g_{20}\left[2\Psi_1(1-\Psi_1)\Psi_2(1-\Psi_2)+\Psi_1{}^2(1-\Psi_2)^2\right]\Big/(1-\Psi_3)^2\\
&+g_{21}\left[2\Psi_1(1-\Psi_1)[1-2\Psi_2(1-\Psi_2)]+\Psi_1{}^2 2\Psi_2(1-\Psi_2)\right]\Big/2\Psi_3(1-\Psi_3)\\
&+g_{22}\left[2\Psi_1(1-\Psi_1)\Psi_2(1-\Psi_2)+\Psi_1{}^2\Psi_2{}^2\right]\Big/\Psi_3{}^2
\end{aligned}$$

Now note that because $\tau_1 < t < \tau_2$,

$$\left(2\Psi_3-1\right)=\left(2\Psi_1-1\right)\left(2\Psi_2-1\right)\text{ i.e. }\Psi_3=\frac{\left(2\Psi_1-1\right)\left(2\Psi_2-1\right)+1}{2}$$

and

$$g_{00}=1-g_{01}-g_{02}-g_{10}-g_{11}-g_{12}-g_{20}-g_{21}-g_{22}$$

Substitution of these into the above formula for $E\left[S(t)\,|\,\Phi\right]$, followed by simplification of each coefficient using MAPLE, led to:

$$
\begin{aligned}
&E\left[S(t)\,|\,\Phi\right] \\
&= \frac{2(\Psi_1-1)(\Psi_2-1)}{1-\Psi_1-\Psi_2+2\Psi_1\Psi_2} + g_{01}\frac{\Psi_1(\Psi_1-1)(2\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)} \\
&+g_{02}\frac{2\Psi_1(\Psi_1-1)(2\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)} \\
&+g_{10}\frac{(2\Psi_1-1)\Psi_2(\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)} + g_{11}\frac{(\Psi_1+\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)} \\
&+g_{12}\frac{(4\Psi_1{}^2\Psi_2-2\Psi_1{}^2+2\Psi_1\Psi_2{}^2-6\Psi_1\Psi_2+2\Psi_1+\Psi_2-\Psi_2{}^2)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)} \\
&+g_{20}\frac{(2\Psi_1-1)2\Psi_2(\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)} \\
&+g_{21}\frac{(2\Psi_1{}^2\Psi_2-\Psi_1{}^2+4\Psi_1\Psi_2{}^2-6\Psi_1\Psi_2+\Psi_1+2\Psi_2-2\Psi_2{}^2)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)} \\
&+g_{22}\frac{2(\Psi_1+\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)}
\end{aligned}
$$

which, after some algebraic manipulation, can be rewritten as:

$$
\begin{aligned}
&E\left[S(t)\,|\,\Phi\right] \\
&=1 \\
&+\frac{\Psi_2(\Psi_2-1)(2\Psi_1-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}\left[g_{10}+g_{11}+g_{12}+2g_{20}+2g_{21}+2g_{22}\right] \\
&+\frac{\Psi_1(\Psi_1-1)(2\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_2\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}\left[g_{01}+g_{11}+g_{21}+2g_{02}+2g_{12}+2g_{22}\right] \\
&-\frac{(2\Psi_1-1)}{1+(2\Psi_1-1)(2\Psi_2-1)}-\frac{(2\Psi_2-1)}{1+(2\Psi_1-1)(2\Psi_2-1)}
\end{aligned}
$$

$$
\begin{aligned}
&=1 \\
&+\frac{\Psi_2(\Psi_2-1)(2\Psi_1-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}E\left[S(\tau_1)\,|\,\Phi\right] \\
&+\frac{\Psi_1(\Psi_1-1)(2\Psi_2-1)}{\left(1-\Psi_1-\Psi_2+2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2-\Psi_1-\Psi_2\right)}E\left[S(\tau_2)\,|\,\Phi\right] \\
&-\frac{(2\Psi_1-1)}{1+(2\Psi_1-1)(2\Psi_2-1)}-\frac{(2\Psi_2-1)}{1+(2\Psi_1-1)(2\Psi_2-1)}
\end{aligned}
$$

Appendix B: Variances of IBD sharing

In chapter 4 (pg 78) we stated that

$$\operatorname{Var}\left(S(\tau_1)\,|\,\Phi\right) = C_1 - C_1^2 + 2\Pr\left(S(\tau_1) = 0\,|\,\Phi\right) \text{ and}$$

$$\operatorname{Var}\left(S(\tau_2)\,|\,\Phi\right) = C_2 - C_2^2 + 2\Pr\left(S(\tau_2) = 0\,|\,\Phi\right).$$

Furthermore,

For $t < \tau_1 < \tau_2$: $\operatorname{Var}\left(S(t)\,|\,\Phi\right) = \left(2\Psi_{t,\tau_1} - 1\right)\left[\operatorname{Var}\left(S(\tau_1)\,|\,\Phi\right) - 0.5\right] + 0.5$.

For $\tau_1 < \tau_2 < t$: $\operatorname{Var}\left(S(t)\,|\,\Phi\right) = \left(2\Psi_{t,\tau_2} - 1\right)\left[\operatorname{Var}\left(S(\tau_2)\,|\,\Phi\right) - 0.5\right] + 0.5$.

For $\tau_1 < t < \tau_2$: $\operatorname{Var}\left(S(t)\,|\,\Phi\right) = E\left(S(t)\,|\,\Phi\right) - \left[E\left(S(t)\,|\,\Phi\right)\right]^2 + 2\Pr\left(S(t) = 2\,|\,\Phi\right)$.

Proof:

Variance of IBD sharing at one of the two disease genes:

$$\begin{aligned}
\operatorname{Var}\left(S(\tau_1)\,|\,\Phi\right) &= \mathrm{E}\left[S^2(\tau_1)\,|\,\Phi\right] - \left\{\mathrm{E}\left[S(\tau_1)\,|\,\Phi\right]\right\}^2 \\
&= \Pr\left(S(\tau_1) = 1\,|\,\Phi\right) + 4\Pr\left(S(\tau_1) = 2\,|\,\Phi\right) - \left[C_1 + 1\right]^2 \\
&= C_1 + 1 + 2\Pr\left(S(\tau_1) = 2\,|\,\Phi\right) - C_1^2 - 2C_1 - 1 \\
&= 2\Pr\left(S(\tau_1) = 2\,|\,\Phi\right) - C_1 - C_1^2
\end{aligned}$$

Noting that

$$\begin{aligned}
E\left[S(\tau_1)\,|\,\Phi\right] &= \Pr\left(S(\tau_1) = 1\,|\,\Phi\right) + 2\Pr\left(S(\tau_1) = 2\,|\,\Phi\right) \\
&= \left[1 - \Pr\left(S(\tau_1) = 0\,|\,\Phi\right) - \Pr\left(S(\tau_1) = 2\,|\,\Phi\right)\right] + 2\Pr\left(S(\tau_1) = 2\,|\,\Phi\right) \\
&= 1 - \Pr\left(S(\tau_1) = 0\,|\,\Phi\right) + \Pr\left(S(\tau_1) = 2\,|\,\Phi\right)
\end{aligned}$$

The variance formula can be re-written as:

$$\begin{aligned}
\operatorname{Var}\left(S(\tau_1)\,|\,\Phi\right) &= 2\left[C_1 + \Pr\left(S(\tau_1) = 0\,|\,\Phi\right)\right] - C_1 - C_1^2 \\
&= C_1 - C_1^2 + 2\Pr\left(S(\tau_1) = 0\,|\,\Phi\right)
\end{aligned}$$

Thus, Var($S(\tau_1)$ |Φ) can be written as a function of C_1 and any one of the three probabilities Pr($S(\tau_1) = i$|Φ) for i = 0,1,2 (formula in terms of Pr($S(\tau_1)$=1|Φ) is not shown but could be similarly derived.)

A similar derivation applies to the derivation of Var($S(\tau_2)$ |Φ).

Variance of IBD sharing at a fully informative marker at location *t*

Case I ($t < \tau_1 < \tau_2$):

$$\begin{aligned}
\operatorname{Var}\left[S(t)\mid\Phi\right] &= \operatorname{E}\left[S^2(t)\mid\Phi\right]-\left\{\operatorname{E}\left[S(t)\mid\Phi\right]\right\}^2 \\
&= \Pr\left(S(t)=1\mid\Phi\right)+4\Pr\left(S(t)=2\mid\Phi\right)-\left[1+(2\Psi-1)C_1\right]^2 \\
&= \Pr\left(S(t)=1\mid S(\tau_1)=0\right)\Pr\left(S(\tau_1)=0\mid\Phi\right) \\
&\quad +\Pr\left(S(t)=1\mid S(\tau_1)=1\right)\Pr\left(S(\tau_1)=1\mid\Phi\right) \\
&\quad +\Pr\left(S(t)=1\mid S(\tau_1)=2\right)\Pr\left(S(\tau_1)=2\mid\Phi\right) \\
&\quad +4\Pr\left(S(t)=2\mid S(\tau_1)=0\right)\Pr\left(S(\tau_1)=0\mid\Phi\right) \\
&\quad +4\Pr\left(S(t)=2\mid S(\tau_1)=1\right)\Pr\left(S(\tau_1)=1\mid\Phi\right) \\
&\quad +4\Pr\left(S(t)=2\mid S(\tau_1)=2\right)\Pr\left(S(\tau_1)=2\mid\Phi\right) \\
&\quad -\left[1+(2\Psi_1-1)C_1\right]^2 \\
&= \left[2\Psi_1(1-\Psi_1)+4(1-\Psi_1)^2\right]\Pr\left(S(\tau_1)=0\mid\Phi\right) \\
&\quad +\left[1-2\Psi_1(1-\Psi_1)+4\Psi_1(1-\Psi_1)\right]\Pr\left(S(\tau_1)=1\mid\Phi\right) \\
&\quad +\left[2\Psi_1(1-\Psi_1)+4{\Psi_1}^2\right]\Pr\left(S(\tau_1)=2\mid\Phi\right) \\
&\quad -\left[1+(2\Psi_1-1)C_1\right]^2
\end{aligned}$$

which after some simplification can be shown to be:

$$\begin{aligned}
\operatorname{Var}\left[S(t)\mid\Phi\right] &= (2\Psi_1-1)\left[C_1-C_1^{\,2}+2\Pr\left(S(\tau_1)=0\mid\Phi\right)-0.5\right]+0.5 \\
&= (2\Psi_1-1)\left[\operatorname{Var}\left[S(\tau_1)\mid\Phi\right]-0.5\right]+0.5
\end{aligned}$$

Case II ($\tau_1 < \tau_2 < t$) follows in a similar way.

For $\tau_1 < t < \tau_2$ (Case III):

$$\begin{aligned}\mathrm{Var}\left[S(t)\,|\,\Phi\right] &= \mathrm{E}\left[S^2(t)\,|\,\Phi\right] - \left[\mathrm{E}[S(t)\,|\,\Phi]\right]^2 \\ &= \Pr\left(S(t)=1\,|\,\Phi\right) + 4\Pr\left(S(t)=2\,|\,\Phi\right) - \left[\mathrm{E}[S(t)\,|\,\Phi]\right]^2 \\ &= \mathrm{E}[S(t)\,|\,\Phi] - \left[\mathrm{E}[S(t)\,|\,\Phi]\right]^2 + 2\Pr\left(S(t)=2\,|\,\Phi\right)\end{aligned}$$

Because of the complicated relationship between Pr($S(t)$=2|Φ) and the nine joint IBD sharing probabilities g_{kl} (k = 0,1,2, l = 0,1,2) when $\tau_1 < t < \tau_2$, we do not further simplify this formula.

Appendix C: Tests for two linked disease genes in a region

Wald Test Details

A testing procedure based on three Wald statistics was proposed in section 7.2.2. Details regarding the three statistics are shown below. The hypothesis H$_{01}$: τ_1 - τ_2 = 0 can be tested using the Wald statistic:

$$T_{wald\_eq\_\tau} = \frac{\left(\hat{\tau}_1 - \hat{\tau}_2\right)^2}{Var(\hat{\tau}_1) + Var(\hat{\tau}_1) - 2Cov(\hat{\tau}_1, \hat{\tau}_2)}$$

A statistic for testing the null hypotheses H$_{02}$: $C_1 = C_2 e^{-.04|\tau_2-\tau_1|}$ may be constructed according to Boos (1992; pg 329, formula $T_{GW\text{II}}$) leading to:

$$T_{wald\_C_1} = h_1(\hat{\delta}_2)^T \left[H_1(\hat{\delta}_2)\mathbf{C\hat{o}v}_R(\hat{\delta}_2)H_1(\hat{\delta}_2)^T\right]^{-1} h_1(\hat{\delta}_2)$$

where h_1 is a vector of equations specifying the null hypothesis as a set of restrictions on the parameters, such that $h_1 = 0$ under the null hypothesis, and H_1 is a matrix of partial derivatives of h_1 with respect to all the parameters, i.e.

$$h_1 = \begin{bmatrix} C_1 - C_2 e^{-.04|\tau_2-\tau_1|} \\ \tau_1 - \tau_{10} \end{bmatrix} \text{ and } H_1 = \frac{\partial h_1}{\partial \delta_2^*} = \begin{bmatrix} 1 & -e^{-.04|\tau_2-\tau_1|} & -.04C_2 e^{-.04|\tau_2-\tau_1|} & .04C_2 e^{-.04|\tau_2-\tau_1|} \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Since here we are interested in testing for excess IBD sharing (beyond that predicted by a one locus model) at the estimated first gene location, $\hat{\tau}_1$, we set $\tau_{10}=\hat{\tau}_1$.

Similarly, for H$_{03}$: $C_2 = C_1 e^{-.04|\tau_2-\tau_1|}$

$$T_{wald\_C_2} = h_2(\hat{\delta}_2)^T \left[H_2(\hat{\delta}_2)\mathbf{C\hat{o}v}_R(\hat{\delta}_2)H_2(\hat{\delta}_2)^T \right]^{-1} h_2(\hat{\delta}_2)$$

where

$h_2 = \begin{bmatrix} C_2 - C_1 e^{-.04|\tau_2-\tau_1|} \\ \tau_2 - \tau_{20} \end{bmatrix}$ and $H_2 = \dfrac{\partial h_2}{\partial \delta_2^*} = \begin{bmatrix} -e^{-.04|\tau_2-\tau_1|} & 1 & .04C_1 e^{-.04|\tau_2-\tau_1|} & -.04C_1 e^{-.04|\tau_2-\tau_1|} \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Here, to test for no effect specifically at our estimated second disease gene location we specify τ_{20} = $\hat{\tau}_2$. Similar test statistics can be formulated for testing $C_1^* = 0$ and $C_2^* = 0$ in the $\delta_2{}^*$-parameterized model.

Under the restrictions imposed by any one of the three component hypotheses, the two-locus mean IBD sharing function reduces to a one-locus mean IBD sharing function, which involves the removal of two free parameters from the model. Thus, the statistics are expected to follow a chi-squared distribution with 2 degrees of freedom. Therefore, a test of the null hypothesis of one disease gene in the region can be based on the statistic

$$T_{wald\_IU} = \min\left(T_{wald\_eq\_\tau}, T_{wald\_C_1}, T_{wald\_C_2}\right)$$

The null hypothesis is rejected if T_{wald_IU} exceeds a critical value from the χ^2_{2df} distribution.

Modified Score Test Details

We may test the null hypothesis of one disease gene in the region, based on the score statistic for the two-locus model parameterized by δ_2 evaluated at $(\hat{C}e^{-.04|\hat{\tau}_1-\hat{\tau}|}, \hat{C}, \hat{\tau}_1, \hat{\tau})$. This is a test of the hypothesis H01: $C_1 = Ce^{-.04|\tau_1 - \tau|}$ for $\tau_1 = \hat{\tau}_1$. Similarly, we can test the hypothesis of no effect of the gene at $\hat{\tau}_2$, i.e. H02: $C_2 = Ce^{-.04|\tau_2 - \tau|}$ for $\tau_2 = \hat{\tau}_2$ using a score statistic evaluated at: $(\hat{C}, \hat{C}e^{-.04|\hat{\tau}_2 - \hat{\tau}|}, \hat{\tau}, \hat{\tau}_2)$. Using the statistic proposed by Rotnizky and Jewell (1993; equation 2.6), the score test statistics to test for deviations from the mean IBD sharing under the null hypothesis, at $\hat{\tau}_1$ and $\hat{\tau}_2$ respectively, would be

$T_{S\_\hat{\tau}_1} = \tilde{U}_2{}' \tilde{\Sigma}^{-1} \tilde{U}_2$, where $\tilde{U}_2 = U_2\left(\hat{C}e^{-.04|\hat{\tau}_1 - \hat{\tau}|}, \hat{C}, \hat{\tau}_1, \hat{\tau}\right)$, and

$T_{S\_\hat{\tau}_2} = \tilde{U}_2{}' \tilde{\Sigma}^{-1} \tilde{U}_2$, where $\tilde{U}_2 = U_2\left(\hat{C}, \hat{C}e^{-.04|\hat{\tau}_2 - \hat{\tau}|}, \hat{\tau}, \hat{\tau}_2\right)$.

U_2 is the score vector for the δ_2-parameterized two-locus model, and Σ is estimated as shown in equation 2.7 or 2.8, i.e. $\tilde{\Sigma} = \sum_{i=1}^{nfam} \tilde{U}_{2i} \tilde{U}_{2i}{}^T$.

Alternatively, $T_{S\_\hat{\tau}_1}$ and $T_{S\_\hat{\tau}_2}$ could be calculated according to the method described by Boos for generalized score tests for hypotheses stated via reparameterization (Boos 1992; formula (7) for T_{GS}, pg. 332)]. We denote these forms of the statistics by $T^B_{S\_\hat{\tau}_1}$ and $T^B_{S\_\hat{\tau}_2}$, respectively. To construct the statistic $T^B_{S\_\hat{\tau}_1}$ we note that the null hypothesis of no additional effect on IBD sharing due to linkage to a gene at $\hat{\tau}_1$, beyond that predicted by a one-locus model, can be written as:

$$\text{H}_{01}\text{:} \begin{pmatrix} C_1 \\ C_2 \\ \tau_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} Ce^{-.04|\hat{\tau}_1 - \tau|} \\ C \\ \hat{\tau}_1 \\ \tau \end{pmatrix} = g(C, \tau) = g(\delta).$$

Let $G(C,\tau) = \partial g(\delta)/\partial\delta$. Also define the matrices *B* and *A* as in section 2.3.2 (equation 2.8), i.e. $B = \sum_{i=1}^{nfam} U_{2i} U_{2i}{}^{T}$, $A = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_2}{\partial \delta_2} \right)' \mathrm{Cov}_2^{-1}(\mathbf{S}^*) \left(\frac{\partial \boldsymbol{\mu}_2}{\partial \delta_2} \right)$. The score statistic in this formulation is given by:

$$T^{B}_{S\_\hat{\tau}_1} = \tilde{U}_2{}' \tilde{\Sigma}^{-1} \tilde{U}_2 \text{ , where } \tilde{\Sigma}^{-1} = \tilde{B}^{-1} - \tilde{B}^{-1}\tilde{A}\tilde{G}(\tilde{G}\tilde{A}\tilde{B}^{-1}\tilde{A}\tilde{G})^{-1}\tilde{G}\tilde{A}\tilde{B}^{-1} .$$

The statistic $T^{B}_{S\_\hat{\tau}_2}$ for testing for additional evidence of linkage at $\hat{\tau}_2$ (beyond the effects of a single disease gene in the region) is constructed in a similar manner. As described in section 7.2.3, the two statistics can then be combined to test the overall hypothesis of no additional evidence of linkage at either $\hat{\tau}_1$ or $\hat{\tau}_2$ (beyond that due to one disease gene in the region) using the statistics:

$$T^{B}_{S\_\max} = \max\left(T^{B}_{S\_\hat{\tau}_1}, T^{B}_{S\_\hat{\tau}_2} \right)$$

and

$$T^{B}_{S\_\mathrm{sum}} = T^{B}_{S\_\hat{\tau}_1} + T^{B}_{S\_\hat{\tau}_2} .$$

Appendix D: Extension of the model for E[*S*(*t*)|Φ] to three or more linked loci

Having derived the expression for E[*S*(*t*)|Φ] for the two-locus scenario, the extension to more linked loci is straightforward. For example, with three linked loci $\tau_1 < \tau_2 < \tau_3$, let $C_1 =$ E(*S*(τ_1)|Φ)-1, $C_2 =$ E(*S*(τ_2)|Φ)-1 and $C_3 =$ E(*S*(τ_3)|Φ)-1.

Then:

for $t<\tau_1<\tau_2<\tau_3$: $$E(S(t) \mid \Phi) = 1 + \left(2\Psi_{t,\tau_1} - 1\right) \cdot C_1$$

for $\tau_1<\tau_2<\tau_3<t$: $$E(S(t) \mid \Phi) = 1 + \left(2\Psi_{t,\tau_3} - 1\right) \cdot C_3$$

for $\tau_1 < t < \tau_2 < \tau_3$:

$$E(S(t)\mid\Phi) = 1 + \frac{\Psi_2\left(\Psi_2 - 1\right)\left(2\Psi_1 - 1\right)}{\left(1 - \Psi_1 - \Psi_2 + 2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2 - \Psi_1 - \Psi_2\right)}C_1 + \frac{\Psi_1\left(\Psi_1 - 1\right)\left(2\Psi_2 - 1\right)}{\left(1 - \Psi_1 - \Psi_2 + 2\Psi_1\Psi_2\right)\left(2\Psi_1\Psi_2 - \Psi_1 - \Psi_2\right)}C_2$$

where $\Psi_1 = \Psi_{t,\tau_1}$ and $\Psi_2 = \Psi_{t,\tau_2}$

and for $\tau_1 < \tau_2 < t < \tau_3$:

$$E(S(t)\mid\Phi) = 1 + \frac{\Psi_3\left(\Psi_3 - 1\right)\left(2\Psi_2 - 1\right)}{\left(1 - \Psi_2 - \Psi_3 + 2\Psi_2\Psi_3\right)\left(2\Psi_2\Psi_3 - \Psi_2 - \Psi_3\right)}C_2 + \frac{\Psi_2\left(\Psi_2 - 1\right)\left(2\Psi_3 - 1\right)}{\left(1 - \Psi_2 - \Psi_3 + 2\Psi_2\Psi_3\right)\left(2\Psi_2\Psi_3 - \Psi_2 - \Psi_3\right)}C_3$$

where $\Psi_2 = \Psi_{t,\tau_2}$ and $\Psi_3 = \Psi_{t,\tau_3}$.

Estimates of τ_1, τ_2, τ_3, and C_1, C_2, and C_3 could, in theory, be obtained by defining $\mu(t)$ = E($S(t)|\Phi$) using the three-locus model shown above and solving estimating equations similar to those for the two-locus model for these six parameters. In practice, however, computational difficulties such as convergence problems are likely to limit the number of linked loci that will be estimable.