THE UNIVERSITY OF CHICAGO

TWO STATISTICAL PROBLEMS IN HUMAN GENETICS:
I. DETECTION OF PEDIGREE ERRORS PRIOR TO GENETIC MAPPING STUDIES
II. IDENTIFICATION OF POLYMORPHISMS THAT EXPLAIN A LINKAGE RESULT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
LEI SUN

CHICAGO, ILLINOIS
AUGUST 2001

to my parents

Contents

List of Figures

List of Tables

Acknowledgments

I am most grateful to my advisor, Mary Sara McPeek, for introducing me to the world of statistical genetics and for her constant guidance, support and encouragement over the past five years. All the work presented here is indebted to her.

I am particularly grateful to the members of my thesis committee, Nancy Cox and Dan Nicolae, for their help and valuable suggestions. I am deeply grateful to all the faculty members of the Department of Statistics for their teachings and support, especially to Xiao-Li Meng. I am also grateful to my colleague, Jian Zhang, for her spending time to explain and discuss statistics with me, to Kenneth Wilder for his generous help in programming and software development, and to Mark Abney and Carole Ober for the collaboration work on the Hutterite data.

I would like to thank a number of fellow students, Peter Bouman, Regina Dolgoarshinnykh, Abby Jager, George Kordzakhia, Wanli Min, James Servidea, Seongjoo Song, Leah Welty and Zhengyuan Zhu, for helping me in many ways.

My warmest thanks go to my dear colleagues, Danielle Harvey and Andrew Strahs, and numerous friends, Yantao Wang, Jia Yu and Gang Zhao, Raluca and Dan Nicolae, Andrei Drăgănescu, Spyros Konstantopoulos, Anda Iamniţchi and Matei Rîpeanu, Bogdan Popescu, Florin Vaida, Ying Zhang and Mircea Pigli, for making the life more beautiful.

Finally, I would like to thank my parents and brother for the love and care they gave me since I was born. Last but not least, Radu, my love, I thank you for every moment that we have shared, for every moment that we are living now, and for every moment that is about to come.

Abstract

We present here two distinct statistical problems in human genetics.

Accurate information on the relationships among individuals is critical for valid genetic mapping studies. Statistical methods for detecting misspecified relationships based on genotype data have been developed mostly for data from sib-pair designs (Chakraborty and Jin 1993a; Göring and Ott 1997; Boehnke and Cox 1997; Ehm and Wagner 1998; Broman and Weber 1998; Olson 1999). We extend the likelihood calculations of Göring and Ott (1997) and Boehnke and Cox (1997) to more general relative pairs for which the identity-by-descent (IBD) process is no longer a Markov process, and we propose the maximized log-likelihood ratio ($MLLR$) test. We also extend the identity-by-state (IBS) test of Ehm and Wagner (1998) to non-sib pairs. The $MLLR$ test has high power but is computationally intensive. The IBS test is simpler, however it ignores information contained in allele frequencies and has low power. To compromise between the two, we then propose two new test statistics, the expected identity by descent ($EIBD$) and the adjusted identity by state ($AIBS$), designed to retain the computational simplicity of IBS, while increasing power by taking into account chance sharing of common alleles. To infer the relationships suggested by the data, we propose a simple method for estimation of pairwise relationships. We describe the implementation of all the methods as freely available software. We perform simulations to compare the power of the methods, and we discuss the applications of our methods to several data sets collected for linkage studies.

The second part of the thesis considers a problem arising in the positional cloning stage of genetic mapping, in which one seeks to identify particular genetic variants affecting susceptibility to complex disease. We assume that a susceptibility locus has been localized, via linkage analysis and fine mapping, to a rather small region of a chromosome, and that many polymorphic sites have been identified and genotyped in that region. A key question of interest is which site or combination of sites in the region influences susceptibility to the trait. We develop here a novel statistical approach to identify the polymorphisms whose genotypes could fully explain the observed linkage to the region. Our approach is based on the observation that if a particular site is the only site in the region that influences the trait, then conditional on the genotypes at that site for the affected relatives, there should be no unexplained over-sharing among the affecteds in the region. The information provided by this analysis is different from that provided by tests of either linkage or association. We focus on the affected sib-pair study design with single nucleotide polymorphism (SNP) data, and we develop test statistics that are variations on the usual allele-sharing methods used in linkage studies. Our method allows for a very general model for how the site influences the trait, including epistasis with unlinked loci, correlated environmental effects within families, and gene-environment interaction. We perform hypothesis tests and derive a confidence set for the true causal polymorphic site, under the assumption that there is only one site in the region influencing the trait. We extend our method to larger sibships and apply it to an $NIDDM1$ data set (Horikawa *et al*. 2000). Both simulation studies and data analysis show that our method can have high power to reject non-causal SNPs, even when they are tightly linked and in strong disequilibrium with the causal SNP. We also discuss the extensions of our method to any set of affected relatives, to any type of causal polymorphism and to multiple tightly-linked causal loci.

Part I

Detection of Pedigree Errors Prior to Genetic Mapping Studies

Chapter 1

Introduction

1.1 Effects of pedigree errors on genetic mapping studies

To map genes or other genetic determinants for a trait or disease, the first step in the process is to collect DNA data on a large number of loci (genetic markers) throughout the genome and perform a coarse genome-wide search via linkage analysis. Linkage analysis looks for regions of genome that are shared by affected relatives, in excess of what is expected under the null hypothesis of no linkage. The excess sharing is evaluated assuming a known pedigree that determines the relationships among the affected individuals and sets the null distribution. However, a sampled pedigree may contain errors. For example, putative full sibs could be half sibs if there is a nonpaternity/nonmaternity, and two hypothesized unrelated individuals could be first cousins if there is a lack of genealogical information. Other sources of pedigree errors include switched samples, duplicated samples, unspecified adoptions and matings between relatives, etc.

Unidentified pedigree errors can have serious consequences for linkage studies. If some individuals were more closely related than the pedigree indicated, then linkage results could be exaggerated and the rate of false positives would be increased. If some individuals were less closely related than the pedigree indicated, then evidence for linkage could be decreased and the power to detect linkage would be reduced. Boehnke and Cox (1997) illustrated the latter case through a sib-pair data set in which 8 putative full-sib pairs were in fact half-sib pairs. They showed that evidence for linkage was increased after the identification and correction of the misspecified relationships.

1.2 From sib pairs to general pairwise relationships

Statistical methods for detecting misspecified relationships based on genotype data have been developed mostly for data from sib-pair designs (Chakraborty and Jin 1993a; Göring and Ott 1997; Boehnke and Cox 1997; Ehm and Wagner 1998; Broman and Weber 1998; Olson 1999). Göring and Ott (1997) and Boehnke and Cox (1997) compute the likelihood of the observed genotype data for full or half-sib pairs under the no-interference model. In the case when a sib pair has a single typed parent, Göring and Ott (1997) also compute the likelihood conditional on the genotype of the parent. The approach of Göring and Ott (1997) is Bayesian. For each putative full-sib pair, they assign prior probabilities to the relationships full-sib, half-sib and unrelated, and they compute the posterior probabilities of the relationships given the data. (The Bayesian approach is discussed in McPeek 2001.) The method of Boehnke and Cox (1997) is analogous to obtaining a point estimate. They first calculate the likelihood of the observed genotype data for a putative full-sib pair, under each of a set of possible relationships, namely, full-sib, half-sib, unrelated and MZ-twin. They then infer the true relationship for the pair by choosing the one that maximizes the likelihood. Although relationship estimation is an essential component of the problem, an additional critical component is the ability to determine whether or not particular relationships are compatible with the observed genotype data. In particular, the hypothesis testing approach is an important one, because the relationship specified by the pedigree has a natural role as the null hypothesis. Ehm and Wagner (1998) propose an approximately normally distributed statistic to test for deviation from a reported relationship of full-sib pair. However, their test loses power because the test statistic ignores the allele-frequency information contained in the genotype data.

In this paper, we first extend the likelihood calculation of Göring and Ott (1997) and Boehnke and Cox (1997), and the work of Ehm and Wagner (1998) to more general relative pairs. We then develop new methods for both detection of pedigree errors and estimation of pairwise relationships. Mapping studies are not restricted to data from sib-pair designs, and families collected may have more than two generations with various pedigree structures, e.g. those in the GAW 11 COGA data (Section 4.1). A more extreme case is the Hutterite data (Section 4.2) in which there is only a single family given by a 13-generation pedigree with about 1600 members in total. To detect pedigree errors in such data, it is necessary to consider the problem of relationship testing for more general relative pairs than sib pairs. We focus on studying pairwise relationships in a pedigree. Compared to a joint analysis, our method may lose some power. However, this pairwise approach results in much simpler implementations and wider applications. In the case of the Hutterite data, it is computationally infeasible to directly analyze the whole pedigree. By breaking a pedigree into pairs, we can also quickly locate the erroneous individuals in the pedigree and propose plausible alternatives for the local structures.

In the next two sections, we introduce mathematical models for the segregation and transmission of chromosomes. We begin with single locus inheritance and then consider multiple loci jointly. The knowledge of how genetic material is passed down through generations is essential to pedigree inference using genetic marker data.

1.3 Single locus inheritance

Genetic material is stored in the chromosomes of every cell. For humans, there are 22 pairs of autosomal chromosomes and a pair of sex-linked chromosomes, with XX for females and XY for males. The genetic material at each chromosomal locus (a particular site on a pair of chromosomes) may be polymorphic, i.e. in different states. The different states are called alleles. The two alleles at a given chromosomal locus constitute a person's genotype at that locus.

Given a locus for an individual, the Mendelian inheritance model specifies that one allele was inherited from the father, and the other was inherited from the mother. Each parent transmits only one of its two alleles to an offspring, each with probability 1/2. Consider each individual's paternally- and maternally-inherited alleles as being random variables. A set of alleles are said to be identical by state (IBS) if they are of the same allelic type, and they are said to be identical by descent (IBD) if they were inherited from the same ancestral allele. (Obviously, IBD implies IBS, ignoring the possibility of mutations.) We illustrate the above description using Figure 1.1.

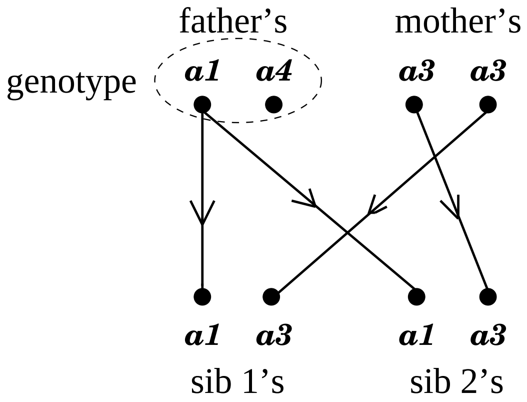

Figure 1.1: Single locus Mendelian inheritance

In this example, at the given locus, the father has alleles $a1$ and $a4$, and the mother has alleles $a3$ and $a3$. The arrows in the graph specify which alleles were transmitted from the parents to the offspring. Among the four alleles of the two sibs, i.e. ($a1$, $a3$, $a1$, $a3$), the two paternally-inherited $a1$ alleles are IBD (also IBS) because they were both inherited from the same $a1$ allele of the father, while the two maternally-inherited $a3$ alleles are IBS (not IBD) because they were inherited from different $a3$ alleles of the mother. In practice, such inheritance information (the arrows in the graph) is often not available. There are two main types of missing information. First, when the genotype of a single individual is collected by the usual method, it is not possible to distinguish the paternally-inherited allele from the maternally-inherited allele. Furthermore, not all loci will be typed for all individuals. In that case, the IBD status may not be unambiguously determined. In the above example, if the parents were not genotyped at the locus, then based on the genotype data for the children, the IBD status could not be determined for either $a1$ allele or $a3$ allele.

The distribution of IBD states for a pair of individuals, $\mathbf{p} = (p_0, p_1, p_2)$, where

$p_0 = P(0$ alleles shared IBD by the pair), $p_1 = P(1$ allele shared IBD by the pair), and $p_2 = P(2$ alleles shared IBD by the pair), can be used to summarize pairwise relationships. For example $\mathbf{p} = (1/4, 1/2, 1/4)$ for a full-sib pair, while $\mathbf{p} = (1/2, 1/2, 0)$ for a half-sib pair. Kinship coefficient Φ is another possible summary of pairwise relationships. Between two individuals i and j, Φ is the probability that a randomly selected allele from individual i and a randomly selected allele from individual j are IBD. Note that Φ is a function of the IBD probabilities, $\Phi = p_1/4 + p_2/2$.

1.4 Multiple locus inheritance

Meiosis is a division process during which an egg or a sperm cell is formed. One important step in meiosis is the recombination or crossover between a pair of chromosomes. During meiosis, each pair of chromosomes first mingle together to exchange genetic material according to a stochastic process, then they separate and transmit only one chromosome strand in each egg or sperm to the next generation. The transmitted chromosome strand is a mixture of the original two. Each child is a product of two independent meioses corresponding to, respectively, the formation of a sperm cell and the formation of an egg cell. Figure 1.2 illustrates the realizations of four independent meioses, M1, M2, M3 and M4, present in a simple pedigree. (In the pedigree graph, a square denotes a male, and a circle denotes a female. Individuals whose parents are not in the graph are called founders, and individuals whose parents are in the graph are called nonfounders.) The illustrated four meiosis are for the two nonfounders, i.e. the two sibs. Figure 1.2 also shows genotype data at three different loci for all the four individuals. Alleles at different loci along one chromosome strand

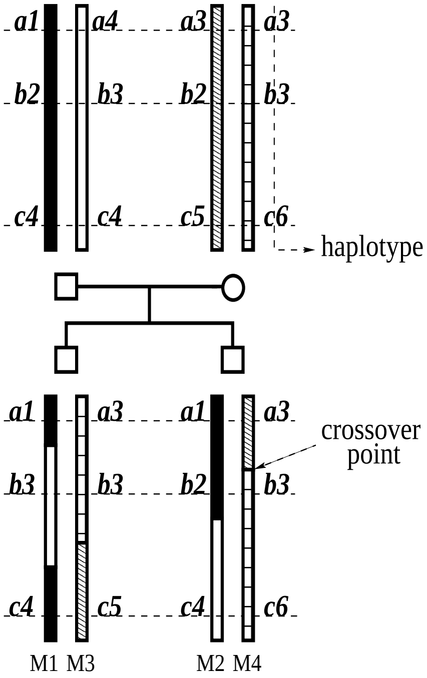

Figure 1.2: Meioses, crossover and haplotype

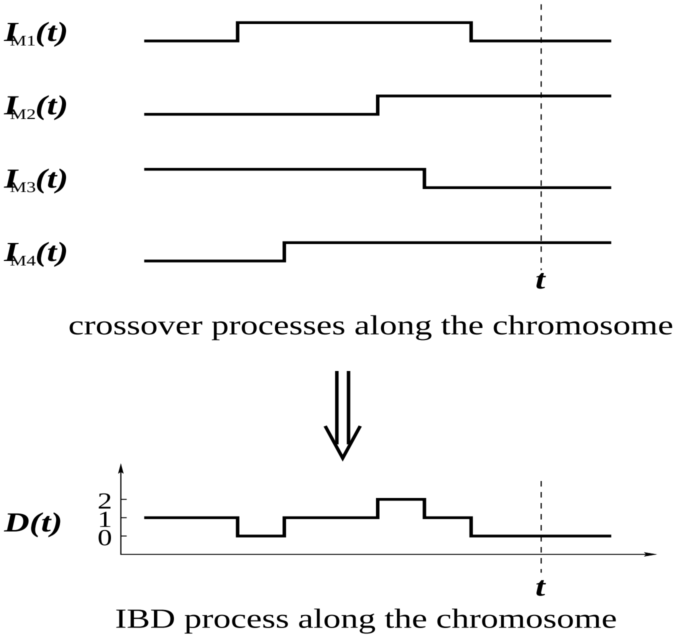
$I_{M1}(t)$
$I_{M2}(t)$
$I_{M3}(t)$
$I_{M4}(t)$
t
crossover processes along the chromosome
D(t)
2
1
0
t
IBD process along the chromosome

Figure 1.3: Crossover and IBD processes

constitute a haplotype, and alleles in the haplotype are said to be in phase. In practice, because it is not possible to distinguish the paternally-inherited allele from the maternally-inherited allele based on the available genotype data at each locus, it is difficult to determine the phase of alleles at different loci.

For a given realization of a meiosis (e.g. M4), a crossover point, as illustrated in Figure 1.2, is defined to be the place where there is a switch between transmission of the paternal DNA and transmission of the maternal DNA. Then we can define a crossover process $\{I(t)\}$, where $I(t)$ is an indicator, at location t along the chromosome, of whether an offspring inherited a given parent's paternal allele ($I(t) = 0$) or maternal allele ($I(t) = 1$). Note that the crossover processes for different chromosomes are assumed to be independent within a meiosis, and the crossover processes for different meioses are also assumed to be independent. For chromosomes M1, M2, M3 and M4 in Figure 1.2, Figure 1.3 depicts the corresponding crossover processes, $\{I_{\mathrm{M1}}(t)\}$, $\{I_{\mathrm{M2}}(t)\}$, $\{I_{\mathrm{M3}}(t)\}$ and $\{I_{\mathrm{M4}}(t)\}$. Consider two loci on a chromosome, t_1 and t_2. Recombination between the two loci denotes the event that the alleles at t_1 and t_2 were inherited from different origins (paternal or maternal alleles), i.e. $I(t_1) = 0$ and $I(t_2) = 1$, or $I(t_1) = 1$ and $I(t_2) = 0$. For example, in Figure 1.2, there is recombination between the two loci a and b on chromosome M1. The recombination fraction θ between two loci is then defined to be the probability that there is recombination between the two loci during a single meiosis. Under some mild assumptions, it can be shown that θ has an upper bound of $1/2$ (see e.g. McPeek 1996). $\theta = 1/2$ between two loci implies that alleles at the loci segregate independently and are unlinked. Linkage is characterized by $\theta < 1/2$. ($\theta = 0$ is called complete or perfect linkage.)

Various stochastic models have been proposed to describe the underlying crossover process. The most commonly used one is called the Haldane no-interference model. Under this model, the transition points of the crossover process $\{I(t)\}$ form a Poisson process with rate 1 per Morgan (a unit of genetic distance). Consequently, $\{I(t)\}$ can be viewed as a continuous-time Markov chain on states 0 and 1, with t corresponding to genetic distance (in units of Morgans) along a chromosome. Morgan is a measure of genetic distance, and it is defined so that the intensity of the process is always 1. Functions that specify the relationships between genetic distance and recombination fraction are called map functions. The Haldane no-interference model implies a specific map function: $\theta = (1 - e^{-2d})/2$, where $d = |t_1 - t_2|$. Although the Haldane no-interference model is commonly used in analysis, the actual data do contain interference (Cobbs 1978; Stam 1979; Foss *et al.* 1993; McPeek and Speed 1995; Zhao *et al.* 1995). It has been shown that the χ^2 model with parameter m provides a reasonable fit to the real data. For humans, Lin and Speed (1996) estimated value of m to be 4. This is discussed further in Section 2.2.4.

Consider a pedigree with f founders and n nonfounders (e.g. the simple pedigree with 2 founders and 2 nonfounders in Figure 1.2). There are $2n$ meioses present in the pedigree, two for each of the nonfounders. For the realization of the k_{th} meiosis, a corresponding crossover process $\{I_k(t)\}$ can be defined. The joint process $\{\mathbf{I}(\mathbf{t})\}$, where $\mathbf{I}(\mathbf{t}) = (I_1(t), I_2(t), ..., I_{2n-1}(t), I_{2n}(t))$, describes the outcomes of all meioses in the pedigree, and it contains complete information on the inheritance pattern in that pedigree. Under the Haldane no-interference model, $\{\mathbf{I}(\mathbf{t})\}$ would be a continuous-time Markov random walk on the vertices of a $2n-$dimensional hypercube (Donnelley 1983). For a particular location t^*, $\mathbf{I}(\mathbf{t}^*) = (I_1(t^*), I_2(t^*), ..., I_{2n-1}(t^*), I_{2n}(t^*))$ is the inheritance vector defined by Lander and Green (1987).

To extend the idea of IBD states at a single locus for a pair of individuals, we can define the IBD process $\{D(t)\}$, with $D(t)$ giving the number of alleles shared IBD by the pair at locus t along a chromosome. Note that the IBD process $\{D(t)\}$ is completely determined by the joint inheritance process $\{\mathbf{I(t)}\}$. Consider the full-sib pair shown in Figure 1.2, with the corresponding crossover processes, $\{I_{\mathrm{M1}}(t)\}$, $\{I_{\mathrm{M2}}(t)\}$, $\{I_{\mathrm{M3}}(t)\}$, $\{I_{\mathrm{M4}}(t)\}$ shown in Figure 1.3. Figure 1.3 (bottom part) illustrates the corresponding IBD process $\{D(t)\}$ that is constructed from the joint inheritance process $\{\mathbf{I(t)}\}$, where $\mathbf{I(t)} = (I_{\mathrm{M1}}(t), I_{\mathrm{M2}}(t), I_{\mathrm{M3}}(t), I_{\mathrm{M4}}(t))$. In practice, neither the inheritance process $\{\mathbf{I(t)}\}$ nor the IBD process $\{D(t)\}$ is observed. In addition to the types of incomplete data described in Section 1.3, the data available are generally discrete observations, i.e. genotype data at a large number of loci for some individuals in a pedigree. Figure 1.2 illustrates a simple case in which only three loci are genotyped for all the individuals. In this example, based only on the genotype data for the individuals, it cannot be discerned whether the two $c4$ alleles of the two sibs are IBS or IBD, and it cannot be discerned whether the $a1$, $b2$ and $c4$ alleles of the father are in phase.

Chapter 2

Likelihood for a Pair of Individuals

2.1 Markov process

To determine whether a relationship is consistent with the observed genotype data for a pair of individuals, it is useful to know the likelihood. To calculate the likelihood, Göring and Ott (1997) Boehnke and Cox (1997) assume that the IBD process is Markov (with an implicit assumption of no interference) and apply a hidden Markov method. However, except in a few simple cases, the IBD process for a pair of individuals does not have the Markov property, even if no interference is assumed. For such a pair, we propose a new process which we call the augmented Markov process, that has the minimum number of states needed to both contain all the information of the IBD process and satisfy the Markov property. In what follows, we first show that the IBD process for a pair is generally not a Markov process, using a specific example of an avuncular pair. We then construct minimal-state augmented Markov processes for a number of types of relative pairs.

2.1.1 IBD process often non-Markov

For a pair of outbred relatives, call them individuals 1 and 2, the IBD process $\{D_t\}$ is a stochastic process giving the number of alleles shared IBD by the pair at locus t along chromosomes. That is,

$$D_t = 1_{g_{11} \equiv g_{21}} + 1_{g_{11} \equiv g_{22}} + 1_{g_{12} \equiv g_{21}} + 1_{g_{12} \equiv g_{22}},$$

where $1_{g_{1i} \equiv g_{2j}}$ is the indicator of the event that allele i of individual 1 and allele j of individual 2 at locus t are IBD, with arbitrary labeling of the two alleles of an individual. For outbred relative pairs, D_t takes values in $\{0, 1, 2\}$ for each t. In order to calculate the likelihood in the cases of full-sib pairs and half-sib pairs, Göring and Ott (1997) and Boehnke and Cox (1997) make the assumption that the IBD process $\{D_t\}$ is Markov. However, as noted by Donnelly (1983) and Feingold (1993), the Markov assumption for $\{D_t\}$ fails to hold in general, although it does hold for a few special cases (MZ-twin, parent-offspring, unrelated, full-sib, half-sib and grandparent-grandchild pairs), when no interference is assumed.

To understand why this is so, first consider an avuncular pair. Let the individuals be labeled 1-6 as in Figure 2.1, with shaded individuals 3 and 6 forming the avuncular pair. The Markov property requires that, conditional on the IBD value D_A for the avuncular pair at a locus A, the IBD values at loci to the right of locus A are independent of the IBD values at loci to the left of locus A. The violation of the Markov property in the avuncular case arises as follows: conditional on the number of alleles shared IBD by individuals 3 and 6 at locus A, if the A allele not transmitted from individual 4 to individual 6 is shared IBD by individuals 3 and 4 (call this event $\mathcal{W}_A$), then the chance is increased that individuals 3 and 6 share an allele IBD at

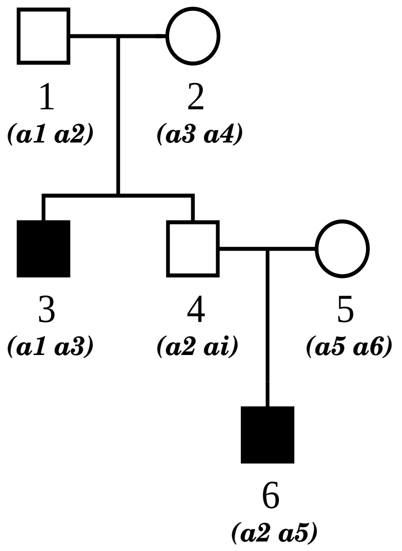

Figure 2.1: Pedigree for an avuncular pair

any other locus linked to A. This induces a positive correlation in sharing at loci linked to A, conditional on IBD sharing at A. By conditioning on the event $\mathcal{W}_A$ or its complement, we show in Appendix A that if locus B is to the right of locus A and locus C is to the left of locus A, both linked to A, then for the avuncular pair 3 and 6,

$$P(D_C = 1|D_A = j, D_B = 1) > P(D_C = 1|D_A = j), \tag{2.1}$$

violating the Markov property. For other cases such as a first-cousin pair in Figure 2.3, the violation of the Markov property for $\{D_t\}$ can be shown using a generalization of the argument for case of a avuncular pair.

2.1.2 Augmented Markov process

In the case where the IBD process $\{D_t\}$ is not Markov, we propose to construct an augmented process $\{A_t\}$ that is Markov under the assumption of no interference and that contains all the information of the IBD process $\{D_t\}$. For an avuncular pair as shown in Figure 2.1, we give the state space for such an augmented Markov process $\{A_t\}$ in Table 2.1. The behavior of a Markov process is determined by its Q-matrix, in which $Q_{ii} = -v_i$ and $Q_{ij} = v_i P_{ij}$, where v_i is the rate at which the process leaves state i, and P_{ij} is the probability that it then goes to state j, i.e. the probability that, given the process is leaving state i, it makes a transition to state j. The Q-matrix of the augmented Markov process $\{A_t\}$ for an avuncular pair is given in Table 2.2, and with the transition probability matrix given in Table 2.3. In Appendix B, we provide results for first-cousin, half-avuncular, half-first-cousin and half-sib-plus-first-cousin pairs.

State label	IBD(3,4) [a]	IBD(3,6) [a]
1	0	0
2	1	0
3	1	1
4	2	1

Table 2.1: State space of $\{A_t\}$ for an avuncular pair

[a] IBD(i, j) is the number of alleles shared IBD by individuals i and j, where individuals are labeled as in Figure 2.1.

Current	Next state entered			
state	1	2	3	4
1	-4	2	2	0
2	2	-5	1	2
3	2	1	-5	2
4	0	2	2	-4

Table 2.2: Q-matrix of $\{A_t\}$ for an avuncular pair

$Q_{ii} = -v_i$, where v_i is the rate (in terms of Morgans) at which the process leaves state i, and $Q_{ij} = v_i P_{ij}$, where P_{ij} is the probability that, given the process is leaving state i, it then makes a transition to state j. States are as labeled in Table 2.1.

Current	State at θ from current state			
state	1	2	3	4
1	ψ^2	$\psi\phi$	$\psi\phi$	ϕ^2
2	$\psi\phi$	$(1-\theta)\psi^2 + \theta\phi^2$	$\theta\psi^2 + (1-\theta)\phi^2$	$\psi\phi$
3	$\psi\phi$	$\theta\psi^2 + (1-\theta)\phi^2$	$(1-\theta)\psi^2 + \theta\phi^2$	$\psi\phi$
4	ϕ^2	$\psi\phi$	$\psi\phi$	ψ^2

Table 2.3: Transition probability matrix of $\{A_t\}$ for an avuncular pair

θ is the recombination fraction between two markers. $\theta = (1-e^{-2t})/2$, where t is the genetic distance (in units of Morgans) between the two markers. $\psi = \theta^2 + (1-\theta)^2$ and $\phi = 1 - \psi = 2\theta(1-\theta)$. States are as labeled in Table 2.1.

In general, under no interference, the augmented Markov process could be chosen to be $\{A'_t\}$, where A'_t is the inheritance vector at location t as defined by Lander and Green (1987) (i.e. $\mathbf{I}(\mathbf{t})$ described in Section 1.4), or $\{A''_t\}$, where A''_t is the equivalence class of inheritance vectors at t defined by Kruglyak *et al.* (1996). $\{A''_t\}$ can be obtained by identifying inheritance vectors that differ only by interchanges of maternal and paternal haplotypes within founders. The use of $\{A'_t\}$ or $\{A''_t\}$ provides an automatic way to construct the augmented Markov process, though these processes contain unnecessary information. (Note that the computational time for the likelihood calculation using the hidden Markov model is proportional to $(\#\text{states})^2$, so it is crucial to reduce the size of the state space of the Markov process used.) For instance, for an avuncular pair, the state space of $\{A'_t\}$ is of size 64, and that for $\{A''_t\}$ is of size 8, while our augmented Markov process $\{A_t\}$ requires only 4 states. Similarly, for a first-cousin pair, the state space of $\{A'_t\}$ is of size 256, that for $\{A''\}$ is of size 16, and that for $\{A_t\}$ is 7. In both cases, our augmented process $\{A_t\}$ has the minimal number of states needed to both contain all the information of the IBD process $\{D_t\}$ and satisfy the Markov property under no-interference assumption (see McPeek 2001). Donnelly (1983) constructs similar minimal-state augmented processes for a number of relationships, such as "m_{th} generation descendant" and "s_{th} cousin t times removed." The problem of implementing an automated method for generating a minimal-state augmented Markov process for any pairwise relationship is treated in McPeek (in press).

2.2 Likelihood calculation using a hidden Markov model

The computation of the likelihood for a pair is strained by many problems such as incomplete inheritance information for the pair, genotype data at many loci (hundreds or thousands markers), possible genotyping errors in the data, etc. To make the likelihood computation faster, we seek a hidden Markov model formulation for the augmented Markov process. We first describe the likelihood calculation for an outbred relative pair, under the assumption of no interference. We then discuss the extensions of the likelihood calculation to allow inbred relationships, to incorporate genotyping errors, and to take into account interference. Finally, we consider a Markov approximation to the IBD process to reduce the computational burden of the likelihood.

2.2.1 Likelihood under the assumption of no interference

First consider outbred relationships. In the case of a full-sib pair, for which the IBD process $\{D_t\}$ is Markov under no interference, Göring and Ott (1997) and Boehnke and Cox (1997) use a hidden Markov model to calculate the probability $P_R(G_1, G_2, ...G_{n_c})$, where n_c is the number of markers on the cth chromosome, G_m is the genotype data for the pair at marker m, and the subscript R denotes that the calculation of the probability is under the assumed relationship R. When the IBD process for a pair is not Markov, the likelihood can still be calculated using the same method, but a hidden Markov model is applied to the augmented Markov process $\{A_t\}$ instead of the IBD process $\{D_t\}$. The calculation, using the Baum forward algorithm (Baum 1972), can be summarized as follows: (i) Define

$$\alpha_1(j) = P_R(A_1 = j), \text{ and}$$
$$\alpha_k(j) = P_R(G_1, G_2, \ldots, G_{k-1}, A_k = j), \text{ for } k > 1.$$

Note that $\{\alpha_1(j) = P_R(A_1 = j) = \pi_j\}$ is just the stationary distribution of the augmented Markov process $\{A_t\}$ for relationship R. The distribution can be inferred from the transition matrix of the process. For instance, the stationary distribution of $\{A_t\}$ for an avuncular pair is $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1/4$, and the stationary distribution of $\{A_t\}$ for a first-cousin pair is $\pi_1 = \pi_2 = \pi_4 = \pi_5 = \pi_6 = \pi_7 = \frac{1}{8}, \pi_3 = \frac{1}{4}$. (ii) Let the recursion formula be

$$\alpha_{k+1}(j) = \sum_i \alpha_k(i) P_R(A_{k+1} = j | A_k = i) P(G_k | A_k = i),$$

where $P_R(A_{k+1} = j | A_k = i)$ is the two-locus transition probability of the augmented Markov process. In the cases of avuncular, first-cousin, half-avuncular and half-first-cousin pairs, the probabilities are given, respectively, in Table 2.3, Table 2.6, Table 2.9 and Table 2.13 in Appendix A. Since the augmented Markov process $\{A_t\}$ contains all the information of the IBD process $\{D_t\}$, and the IBD status is sufficient to calculate the conditional probability of genotype data, we obtain $P(G_k | A_k = i) = P(G_k | D_k =$ the IBD status associated with state i of A). (In fact, given the IBD status at a marker for a pair, the probability of genotype data does not depend on the assumed relationship for the pair.) Thus, in the case of a outbred pair, the calculation of $P(G_k | A_k = i)$ can be reduced to the calculation of $P(G_k | D_k = j)$ that is given in Thompson (1975). (iii) The following summation gives the probability of genotype data on the cth chromosome:

$$P_R(G_1, G_2, \ldots, G_{nc}) = \sum_j \alpha_{nc}(j) P(G_{nc} | D_{nc} = j).$$

(iv) Because meioses on different chromosomes are independent, a multiplication over all chromosomes gives the likelihood for genotype data throughout the genome under relationship R,

$$P_R(\text{genome-screen data}) = \Pi_c P_R(G_1, G_2, \ldots, G_{n_c}).$$

2.2.2 Inbred relative pairs

We now consider inbred relationships. In the case of an inbred relative pair, instead of 3 IBD states (0, 1 or 2 alleles shared IBD), there are now 9 condensed identity states (Jacquard 1970). Let the (unordered) genotype of individual 1 be (G_1, G_2) and let the (unordered) genotype of individual 2 be (G_3, G_4). Jacquard (1970) depicts each of the condensed identity states by a graph with four nodes, each representing one of G_1, G_2, G_3, G_4, with an edge present between G_i and G_j if and only if G_i and G_j are IBD. See Figure 2.2. States S_7, S_8, and S_9 correspond to outbred states of 2, 1, and 0 alleles shared IBD, respectively. The other 6 identity states involve inbreeding in one or both individuals. In principle, an augmented Markov process could be derived, e.g. the process $\{A_t''\}$ as described in Section 2.1.2, and a hidden Markov method could be applied to $\{A_t''\}$ to calculate the likelihood. To perform this calculation, we need the distribution, given in Jacquard (1970), of genotype conditional on the condensed identity state. For inbred pairs, the computational burden of the likelihood calculation is generally high. The reason is that the algorithm used in our likelihood calculation scales quadratically in the number of states of the Markov chain and linearly in the number of markers. The size of the state space of $\{A_t''\}$ for a pair of individuals is 2^{2n-f}, where n is the number of nonfounders and f is the number

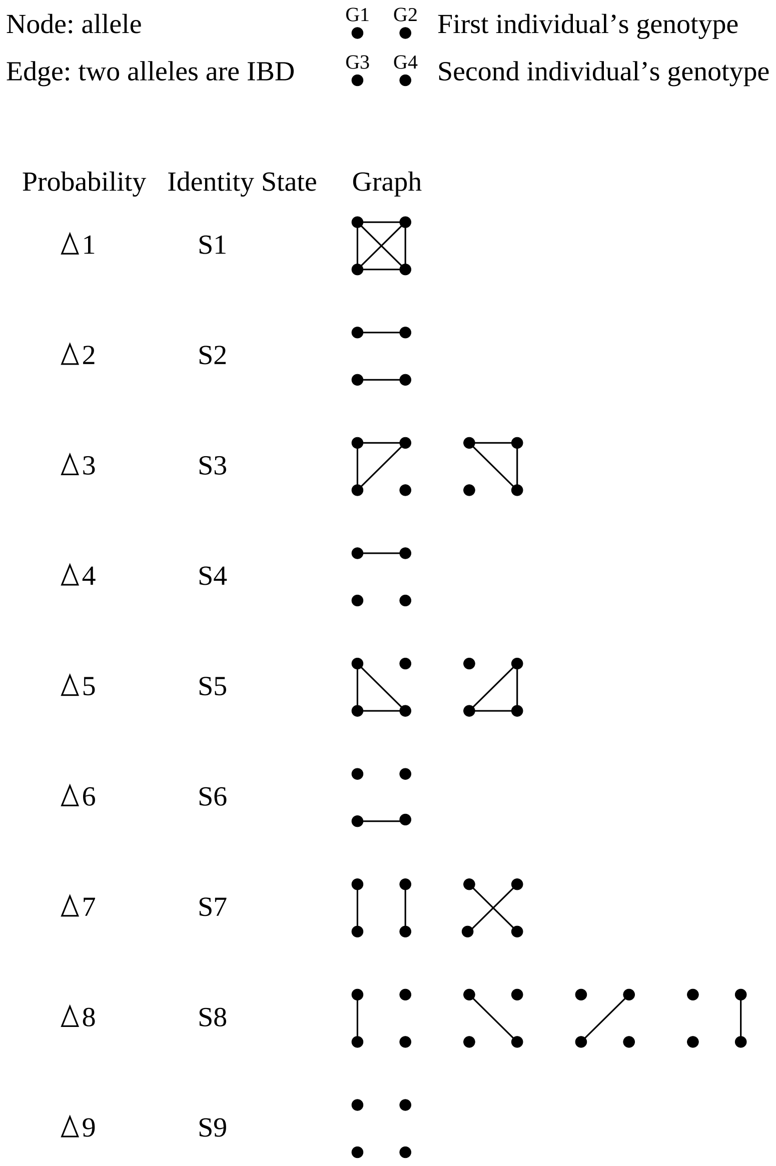

Figure 2.2: Nine condensed identity states

of founders in the pedigree graph that connects the two individuals. $(2n - f)$ tends to be large for an inbred relative pair because they tend to have a larger number of internal edges in their pedigree graph than do a outbred relative pair. For instance, $(2n-f) = 10$ for siblings who are offspring from a second-cousin mating. For pairs in the Hutterite pedigree, the likelihood calculation becomes computationally infeasible.

2.2.3 Likelihood in the presence of genotyping errors

Genotype data may be contaminated during various stages in the process of data collection. As a result, some observed genotype data may not represent the true underlying genetic information and may contain errors. For a MZ-twin pair or a parent-offspring pair, a single genotyping error may result in 0 alleles shared IBS at a particular marker, which leads to zero likelihood under the models for these relationships. To make the likelihood calculation more robust to the presence of genotyping errors for these two models, one can adjust the conditional distribution of genotype data given the IBD status at the k_{th} marker, $c_i = P(G_k|D_k = i), i \in \{0, 1, 2\}$, so that c_1 and c_2 will always be non-zero (Broman and Weber 1998) . Consider the random-genotype-error model in which we assume that genotype errors occur independently with the same probability ϵ across different alleles and different markers. Then, in the likelihood calculation, we could replace c_i, $i \in \{0, 1, 2\}$ with

$$\begin{aligned} {c_o}^* &= c_0, \\ {c_1}^* &= (1-\epsilon)^2 c_1 + (1 - (1-\epsilon)^2) c_0, \\ {c_2}^* &= (1-\epsilon)^2 c_2 + (1 - (1-\epsilon)^2) c_0. \end{aligned}$$

In practice, the random-genotype-error model assumption may not hold, and the true error rate is often unknown. Based on the results of Epstein *et al.* (2000), it appears that the random-genotype-error model works adequately for detection of pedigree errors from genome-screen data, and the likelihood approach is robust to moderate misspecification of the genotype-error rate, as long as the assumed error rate is not zero for MZ-twin and parent-offspring pairs.

2.2.4 Likelihood taking into account interference

The likelihood calculation under the no-interference assumption can be extended to certain models allowing for interference, e.g. the χ^2 model. Under the χ^2 interference model, crossover points occur following a stationary renewal process with inter-arrivals distributed as a mixture of χ^2 random variables. The χ^2 model can be viewed as a hidden Markov model. By applying the Baum algorithm, the likelihood can be obtained. To see this, note that for the χ^2 model with parameter m, the crossover process on four strands can be obtained by first constructing a Poisson process with rate $2(m + 1)$ in terms of genetic distance. Start at one end of the chromosome and label the first point of the Poisson process X_1, where X_1 is chosen uniformly at random from integers $\{0, 1, ..., m\}$. For all $i > 1$, label the ith point of the process (counting in order from the end of the chromosome) $X_i = X_1 + i \pmod{m + 1}$. Then every point with label 0 is a crossover point for the four-strand process. To obtain the single-strand crossover process, independently thin each point of the four-strand crossover process with chance $1/2$. Consider a single strand inherited by an offspring from its parent. For a given chromosomal location t, define $Z(t) = (X(t), Y(t))$, where $X(t) = X_i$ for $C_i \leq t < C_{i+1}$, C_i being the position of the i_{th} point of the original Poisson$(2(m+1))$ process, and $Y(t) = 1$ if the offspring inherited the given parent's paternal DNA and 0 if the offspring inherited the given parent's maternal DNA at location t. Then $Z(t) = (X(t), Y(t))$ is a Markov chain with P(Next state is $(x2, y2)|$ Current state is $(x1, y1)) = P_{(x1,y1),(x2,y2)}$, where $P_{(i,j),(i+1,j)} = 1$ for $i \in \{0, 1, ..., m-1\}$, $j \in \{0, 1\}$, $P_{(m,j),(0,k)} = 1/2$ for $j, k \in \{0, 1\}$, and all other entries are 0. The leaving rate for each state is $2(m+1)$. The observed data give partial information on $Y(t)$ only. Thus, the Markov chain $Z(t) = (X(t), Y(t))$ is hidden. Now consider a pair of individuals in a pedigree. Define such a hidden Markov chain for each meiosis in the pedigree, and consider the product Markov chain $(Z_1(t), ..., Z_n(t)) = (X_1(t), Y_1(t), ..., X_n(t), Y_n(t))$, where $Z_i(t) = (X_i(t), Y_i(t))$ is the Markov chain for the ith meiosis in the pedigree, and there are n meioses in total. Since the meioses are independent, the transition probability matrix is the n-fold Kronecker product of the transition probability matrix for a single meiosis. If genotype data for the pair of individuals is observed, then in principle the Baum algorithm can be applied to the product Markov chain to calculate the likelihood. As in Kruglyak *et al.* (1996), a reduction in dimensionality could be achieved by identifying states that differ by one or more interchanges of founders' paternally and maternally inherited haplotypes.

2.2.5 Likelihood approximation

To reduce the computational burden of the likelihood, one strategy is to approximate the IBD process $\{D_t\}$ by a Markov process $\{B_t\}$, with the probability

$P_R(D_{m_2} = j|D_{m_1} = i)$ used as the transition probability for $\{B_t\}$, where m_1 and m_2 label adjacent markers. That is, $\{B_t\}$ is a Markov chain on a set of markers, with $P_R(B_{m_1} = i) = P_R(D_{m_1} = i)$ for all markers $m1$, and $P_R(B_{m_2} = j|B_{m_1} = i) = P_R(D_{m_2} = j|D_{m_1} = i)$ for all adjacent markers m_1 and m_2. This would eliminate the need to construct augmented Markov processes for general pairwise relationships. The likelihood would then be calculated using $\{B_t\}$ as the hidden Markov chain. Algorithms for calculating $P_R(D_{m_2} = j|D_{m_1} = i)$ for some outbred relationships are discussed in Denniston (1975), Thompson (1988) and Tiwari and Elston (1999). More generally, the probabilities can be determined from $\{A'_t\}$ or $\{A''_t\}$. Explicit formulae are derived by Bishop and Williamson (1990) for full-sib, half-sib, parent-offspring, grandparent-grandchild, avuncular and first-cousin pairs. We give the probabilities in Appendix A for half-avuncular, half-first-cousin and half-sib-plus-first-cousin pairs. Note that MZ-twin, parent-offspring and unrelated pairs have degenerate IBD processes, i.e. these processes are constant everywhere: $P(D_{m_2} = 2|D_{m_1} = 2) \equiv 1$, $P(D_{m_2} = 1|D_{m_1} = 1) \equiv 1$ and $P(D_{m_2} = 0|D_{m_1} = 0) \equiv 1$ for the three cases respectively.

For avuncular, first-cousin, half-avuncular, half-first-cousin and half-sib-plus-first-cousin pairs, we perform simulation to compare the likelihood calculated using the augmented Markov process (correct likelihood) with the incorrect likelihood obtained using the Markov approximation to the IBD process as describe above. Our simulation consists of 10^5 replicates for each of the five relationships considered, using a realistic set of 300 microsatellite markers from the Marshfield map (Broman *et al.* 1998). These markers are unevenly spaced along the genome (excluding sex chromosomes), and they have different allele frequency distributions. Out of the 10^5

replicates, the maximum relative errors of the likelihood using the Markov approximation are, respectively, 1.2×10^{-4}, 1.7×10^{-4}, 6.1×10^{-4}, 8.1×10^{-4} and 6.9×10^{-4} for the five relationships. This suggests that, at least for the cases considered, the Markov approximation to the likelihood is adequate. The theoretical basis of this result for general pairwise relationships needs to be examined.

2.3 Appendix A

To see that the avuncular, first-cousin, half-sib, half-first-cousin, half-sib-plus-first-cousin IBD processes are not Markov, it would suffice to provide a counterexample in each case. However, in order to understand the augmented Markov processes we introduce, it is necessary to understand the nature of the violation of the Markov property, which we now describe in more detail.

First consider the avuncular case. Suppose the genotypes at locus A are as given in Figure 2.1, where the maternally inherited allele of individual 4, a_i, is either a_3 or a_4. Here, the avuncular pair of individuals 3 and 6 share 0 alleles IBD at locus A. Consider a nearby locus B to the right of locus A, linked to A. We are interested the distribution of the number of alleles shared IBD by individuals 3 and 6 at locus B conditional on the genotype information at locus A. We make the relatively weak assumption that the crossover process is a regular stationary point process, with chiasma interference permitted but with no chromatid interference, that is, the choices of chromatid strands for different crossovers are independent and uniform. In the example shown in Figure 2.1, if a_i is a_3, i.e. if the allele not transmitted from 4 to 6 is shared IBD by individuals 3 and 4, then the event of 1 allele shared IBD by the avuncular pair at locus B may be achieved by a single crossover in any one of 3 meioses: the meioses involving transmission of genetic material from individual 1 to individual 3, 1 to 4, or 4 to 6. If a_i is a_4, i.e. if the allele not transmitted from 4 to 6 is not shared IBD between individuals 3 and 4, then the same event of 1 allele shared IBD by the avuncular pair at locus B may be achieved by a single crossover in either of 2 meioses: 1 to 3, or 1 to 4. Thus, the instantaneous rate of transition at A from 0 IBD to 1 IBD for the avuncular pair is 3 if $a_i = a_3$ and 2 if $a_i = a_4$. This suggests that in the example shown in Figure 2.1, if we condition on all the allele information at locus A, then the distribution of the number of alleles shared IBD by the avuncular pair at B depends on a_i, that is, it depends on whether the A allele not transmitted from individual 4 to individual 6 is shared IBD between individuals 3 and 4. In fact, letting D_A and D_B be the number of alleles shared IBD by individuals 3 and 6 at loci A and B, letting $\mathcal{W}_A$ denote the event that the A allele not transmitted from individual 4 to individual 6 is shared IBD between individuals 3 and 4, and letting $\mathcal{W}_A^c$ denote the complementary event to $\mathcal{W}_A$, we have that $P(D_B = 1|D_A = j, \mathcal{W}_A) > P(D_B = 1|D_A = j, \mathcal{W}_A^c)$ when A and B are linked. This inequality can be deduced from the transition probabilities given in Table 2.3, along with the fact that $2\theta(1-\theta) < \theta^2 + (1-\theta)^2$ for $0 \le \theta < 1/2$. Immediate consequences of this inequality are

(i) $P(D_B = 1|D_A = j, \mathcal{W}_A) > P(D_B = 1|D_A = j)$,

(ii) $P(D_B = 0|D_A = j, \mathcal{W}_A) < P(D_B = 0|D_A = j, \mathcal{W}_A^c)$,

(iii) $P(D_B = 0|D_A = j) > P(D_B = 0|D_A = j, \mathcal{W}_A)$.

Using (i) and (iii), we have that

$$\begin{aligned}
& P(\mathcal{W}_A|D_A = j, D_B = 1) \\
= \;& P(\mathcal{W}_A, D_B = 1|D_A = j)/P(D_B = 1|D_A = j) \\
= \;& P(D_B = 1|D_A = j, \mathcal{W}_A)P(\mathcal{W}_A|D_A = j)/P(D_B = 1|D_A = j) \\
> \;& P(D_B = 0|D_A = j, \mathcal{W}_A)P(\mathcal{W}_A|D_A = j)/P(D_B = 0|D_A = j) \\
= \;& P(\mathcal{W}_A|D_B = 0, D_A = j).
\end{aligned}$$

This implies (iv) $P(\mathcal{W}_A|D_A = j, D_B = 1) > P(\mathcal{W}_A|D_A = j)$. Thus, conditional on the number of alleles shared IBD by the avuncular pair at locus A, the probability that $\mathcal{W}_A$ occurs is increased if the IBD sharing by the avuncular pair is 1 at a nearby locus B. Suppose B is to the right of A, and let C be another nearby locus to the left of A. Then using (iv) and the fact that $(\mathcal{W}, D)$ is Markov, we have that

$$\begin{aligned}
& P(D_C = 1|D_A = j, D_B = 1) \\
= \;& P(D_C = 1|D_A = j, \mathcal{W}_A, D_B = 1)P(\mathcal{W}_A|D_A = j, D_B = 1) \\
& +P(D_C = 1|D_A = j, \mathcal{W}_A^c, D_B = 1)P(\mathcal{W}_A^c|D_A = j, D_B = 1) \\
= \;& P(D_C = 1|D_A = j, \mathcal{W}_A)P(\mathcal{W}_A|D_A = j, D_B = 1) \\
& +P(D_C = 1|D_A = j, \mathcal{W}_A^c)[1 - P(\mathcal{W}_A|D_A = j, D_B = 1)] \\
= \;& P(D_C = 1|D_A = j, \mathcal{W}_A^c) + [P(D_C = 1|D_A = j, \mathcal{W}_A) \\
& -P(D_C = 1|D_A = j, \mathcal{W}_A^c)]P(\mathcal{W}_A|D_A = j, D_B = 1) \\
> \;& P(D_C = 1|D_A = j, \mathcal{W}_A^c) + [P(D_C = 1|D_A = j, \mathcal{W}_A) \\
& -P(D_C = 1|D_A = j, \mathcal{W}_A^c)]P(\mathcal{W}_A|D_A = j) \\
= \;& P(D_C = 1|D_A = j).
\end{aligned}$$

This shows inequality (2.1) in Section , in violation of the Markov property. The proof for other cases such as a first-cousin pair follows, using a generalization of this argument.

2.4 Appendix B

In this appendix, we derive the minimal-state augmented Markov processes $\{A_t\}$ for first-cousin, half-avuncular, half-first-cousin and half-sib-plus-first-cousin pairs. For each relative pair, we first illustrate the relevant pedigree structure using a graph in which the shaded two individuals have the relationship of interest. For the first three relative pairs, we then give the state space, Q-matrix and transition probability matrix of the augmented Markov process $\{A_t\}$ for that pair. We also calculate the transition probabilities of the IBD processes $\{D_t\}$ for these pairs. These probabilities are used in Section 2.2.5 and are need for Section 3.1. Note that the augmented Markov process for a half-sib-plus-first-cousin pair can be viewed as a combination of two independent processes: the Markov IBD process for a half-sib pair and the augmented Markov process for a first-cousin pair. Thus, we do not show the state space, Q-matrix and transition probability matrix of the augmented Markov process for this relationship. We define the notation used in this appendix: $\{A_t\}$ is the minimal-state augmented Markov process, $\{D_t\}$ is the IBD process, θ is the recombination fraction between two loci, t is genetic distance (in units of Morgans) between the loci, $\theta = (1 - e^{-2t})/2$, $\psi = \theta^2 + (1 - \theta)^2$, $\phi = 1 - \psi$, $Q_{ii} = -v_i$, where v_i is the rate (in terms of Morgans) at which the process leaves state i, and $Q_{ij} = v_i P_{ij}$, where P_{ij} is the probability that the process then makes a transition to state j.

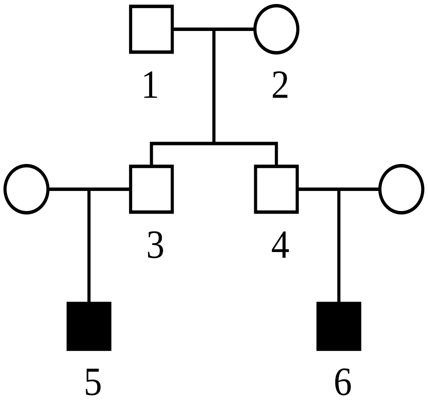

Figure 2.3: Pedigree for a first-cousin pair

State label	IBD(3,4) [a]	IBD(5,6) [a]	G(5,6) [b]
1	0	0	0
2	0	0	1
3	1	0	0
4	1	0	1
5	1	1	1
6	2	0	0
7	2	1	1

Table 2.4: State space of $\{A_t\}$for a first-cousin pair

[a] IBD(i, j) is the number of alleles shared IBD by individuals i and j.

[b] G(5,6) is the indicator of the event that the allele inherited by individual 5 from individual 3 and the allele inherited by individual 6 from individual 4 are both descended from individual 1 or both descended from individual 2, i.e. that they are both from the same grandparent. Individuals are labeled as in Figure 2.3.

Current	Next state entered						
state	1	2	3	4	5	6	7
1	-6	2	4	0	0	0	0
2	2	-6	0	2	2	0	0
3	2	0	-6	1	1	2	0
4	0	2	2	-6	0	0	2
5	0	2	2	0	-6	0	2
6	0	0	4	0	0	-6	2
7	0	0	0	2	2	2	-6

Table 2.5: Q-matrix of $\{A_t\}$ for a first-cousin pair

Current	State at θ from current state						
state	1	2	3	4	5	6	7
1	ψ^3	$\psi^2\phi$	$2\psi^2\phi$	$\psi\phi^2$	$\psi\phi^2$	$\psi\phi^2$	ϕ^3
2	$\psi^2\phi$	ψ^3	$2\psi\phi^2$	$\psi^2\phi$	$\psi^2\phi$	ϕ^3	$\psi\phi^2$
3	$\psi^2\phi$	$\psi\phi^2$	$\psi^3+\psi\phi^2$	$\theta(1-\theta)(\psi^2+\phi^2)$	$\theta(1-\theta)(\psi^2+\phi^2)$	$\psi^2\phi$	$\psi\phi^2$
4	$\psi\phi^2$	$\psi^2\phi$	$\psi^2\phi+\phi^3$	$(1-\theta)^2\psi^2+\theta^2\phi^2$	$\theta^2\psi^2+(1-\theta)^2\phi^2$	$\psi\phi^2$	$\psi^2\phi$
5	$\psi\phi^2$	$\psi^2\phi$	$\psi^2\phi+\phi^3$	$\theta^2\psi^2+(1-\theta)^2\phi^2$	$(1-\theta)^2\psi^2+\theta^2\phi^2$	$\psi\phi^2$	$\psi^2\phi$
6	$\psi\phi^2$	ϕ^3	$2\psi^2\phi$	$\psi\phi^2$	$\psi\phi^2$	ψ^3	$\psi^2\phi$
7	ϕ^3	$\psi\phi^2$	$2\psi\phi^2$	$\psi^2\phi$	$\psi^2\phi$	$\psi^2\phi$	ψ^3

Table 2.6: Transition probability matrix of $\{A_t\}$ for a first-cousin pair

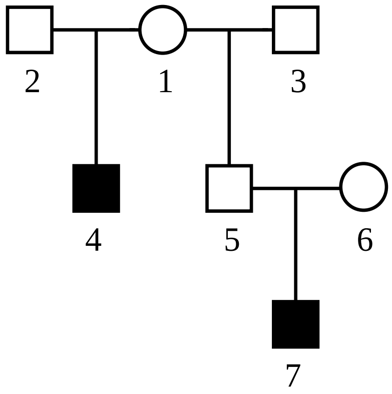

Figure 2.4: Pedigree for a half-avuncular pair

State label	IBD(4,5) [a]	G(7) [b]	IBD(4,7) [a]
1	0	0	0
2	0	1	0
3	1	0	0
4	1	1	1

Table 2.7: State space of $\{A_t\}$for a half-avuncular pair

[a] IBD(i, j) is the number of alleles shared IBD by individuals i and j.

[b] G(7) is the indicator of the event that the allele inherited by individual 7 from individual 5 is descended from individual 1. Individuals are labeled as in Figure 2.4.

Current	Next state entered			
state	1	2	3	4
1	-3	1	2	0
2	1	-3	0	2
3	2	0	-3	1
4	0	2	1	-3

Table 2.8: Q-matrix of $\{A_t\}$ for a half-avuncular pair

Current	State at θ from current state			
state	1	2	3	4
1	$(1-\theta)\psi$	$\theta\psi$	$(1-\theta)\phi$	$\theta\phi$
2	$\theta\psi$	$(1-\theta)\psi$	$\theta\phi$	$(1-\theta)\phi$
3	$(1-\theta)\phi$	$\theta\phi$	$(1-\theta)\psi$	$\theta\psi$
4	$\theta\phi$	$(1-\theta)\phi$	$\theta\psi$	$(1-\theta)\psi$

Table 2.9: Transition probability matrix of $\{A_t\}$ for a half-avuncular pair

Current	IBD state at θ from current IBD state	
IBD state	0	1
0	$\frac{2}{3}+\frac{1}{3}P_{11}$	$\frac{1}{3}(1-P_{11})$
1	$1-P_{11}$	P_{11}

Table 2.10: Transition probability matrix of $\{D_t\}$ for a half-avuncular pair

$$P_{11} = (1-\theta)^3 + \theta^2(1-\theta).$$

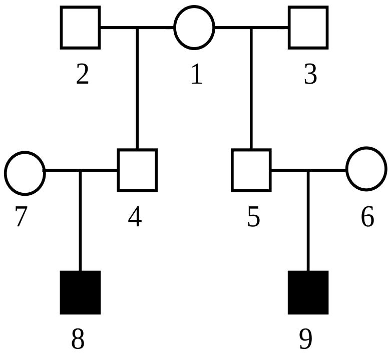

Figure 2.5: Pedigree for a half-first-cousin pair

State label	IBD(4,5) [a]	G(8,9) [b]	IBD(8,9) [a]
1	0	0	0
2	0	1	0
3	0	2	0
4	1	0	0
5	1	1	0
6	1	2	1

Table 2.11: State space of $\{A_t\}$for a half-first-cousin pair

[a] IBD(i, j) is the number of alleles shared IBD by individuals i and j.

[b] G(8,9) is the sum of two indicator functions: the indicator of the event that the allele inherited by individual 8 from individual 4 is descended from individual 1 and the indicator of the event that the allele inherited by individual 9 from individual 5 is descended from individual 1. That is, if they are both from individual 1, G(8,9) = 2, if none are from individual 1, G(8,9) = 0, otherwise G(8,9) = 1. Individuals are labeled as in Figure 2.5.

Current	Next state entered					
state	1	2	3	4	5	6
1	-4	2	0	2	0	0
2	1	-4	1	0	2	0
3	0	2	-4	0	0	2
4	2	0	0	-4	2	0
5	0	2	0	1	-4	1
6	0	0	2	0	2	-4

Table 2.12: Q-matrix of $\{A_t\}$ for a half-first-cousin pair

Current	State at θ from current state					
state	1	2	3	4	5	6
1	$(1-\theta)^2\psi$	$\psi\phi$	$\theta^2\psi$	$(1-\theta)^2\phi$	ϕ^2	$\theta^2\phi$
2	$\theta(1-\theta)\psi$	ψ^2	$\theta(1-\theta)\psi$	$\theta(1-\theta)\phi$	$\psi\phi$	$\theta(1-\theta)\phi$
3	$\theta^2\psi$	$\psi\phi$	$(1-\theta)^2\psi$	$\theta^2\phi$	ϕ^2	$(1-\theta)^2\phi$
4	$(1-\theta)^2\phi$	ϕ^2	$\theta^2\phi$	$(1-\theta)^2\psi$	$\psi\phi$	$\theta^2\psi$
5	$\theta(1-\theta)\phi$	$\psi\phi$	$\theta(1-\theta)\phi$	$\theta(1-\theta)\psi$	ψ^2	$\theta(1-\theta)\psi$
6	$\theta^2\phi$	ϕ^2	$(1-\theta)^2\phi$	$\theta^2\psi$	$\psi\phi$	$(1-\theta)^2\psi$

Table 2.13: Transition probability matrix of $\{A_t\}$ for a half-first-cousin pair

Current	IBD state at θ from current IBD state	
IBD state	0	1
0	$\frac{6}{7}+\frac{1}{7}P_{11}$	$\frac{1}{7}(1-P_{11})$
1	$1-P_{11}$	P_{11}

Table 2.14: Transition probability matrix of $\{D_t\}$ for a half-first-cousin pair $P_{11}=(1-\theta)^4+\theta^2(1-\theta)^2$.

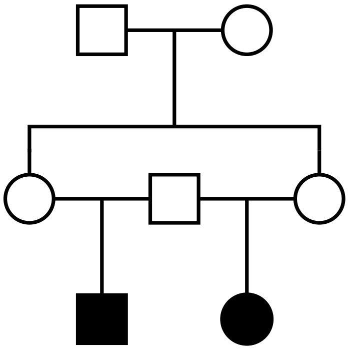

Figure 2.6: Pedigree for a half-sib-plus-first-cousin pair

Current	IBD state at θ from current IBD state		
IBD state	0	1	2
0	ψP_{00}	$1-\psi P_{00}-\phi P_{01}$	ϕP_{01}
1	$\frac{2}{3}\phi P_{00}+\frac{1}{3}\psi P_{10}$	$1-(\frac{2}{3}\phi P_{00}+\frac{1}{3}\psi P_{10})-(\frac{2}{3}\psi P_{01}+\frac{1}{3}\phi P_{11})$	$\frac{2}{3}\psi P_{01}+\frac{1}{3}\phi P_{11}$
2	ϕP_{10}	$1-\phi P_{10}-\psi P_{11}$	ψP_{11}

Table 2.15: Transition probability matrix of $\{D_t\}$ for a half-sib-plus-first-cousin pair

$P_{11}=(1-\theta)^4+\theta^2(1-\theta)^2+1/2\theta^2$, $P_{10}=1-P_{11}$, $P_{00}=1-P_{01}$, $P_{01}=1/3(1-P_{11})$, where P_{ij}, $i,j\in\{0,1\}$, are the transition probabilities of IBD process $\{D_t\}$ for a first-cousin pair.

Chapter 3

Error Detection and Relationship Estimation

3.1 Tests for detection of misspecified relationships

We first consider a likelihood approach and define a test statistic that is the maximized log-likelihood ratio ($MLLR$). Although the $MLLR$ test has high power, it is computationally intensive. As alternatives, we propose new tests based on the identity by state (IBS), expected identity by descent ($EIBD$) and adjusted identity by state ($AIBS$) statistics. The test statistic IBS is a modification of that of Ehm and Wagner (1998). Compared to the $MLLR$ test, the IBS test is easy to perform, but it loses power by ignoring the information contained in allele frequencies. $EIBD$ and $AIBS$ are designed to correct for chance sharing of common alleles, while retaining the computational simplicity of IBS. We develop these tests in the context of outbred relationships. In Appendix C, we discuss their extensions to the case of an inbred relative pair.

3.1.1 The $MLLR$ test

Let L_R be the likelihood of genotype data for a pair of individuals with an assumed relationship R, that is,

$$L_R = P(\text{genotype data for a pair} \mid \text{relationship } R).$$

Given the Markov or augmented Markov process for a relationship, the likelihood can be calculated as described in the previous chapter. For a pair of individuals with null relationship R_o, in order to test R_o against a specific alternative relationship R_a (e.g. H_o: the full-sib relationship; H_a: the half-sib relationship), we can calculate the likelihoods L_{R_o} and L_{R_a} and consider the likelihood ratio test. The test statistic could be the log-likelihood ratio (LLR)

$$LLR = \log(L_{R_a}) - \log(L_{R_o}).$$

However, in practice, we often do not have a specific alternative relationship in mind. For instance, putative full sibs could be half sibs due to nonpaternity/nonmaternity, unrelated individuals due to adoption or MZ twins (monozygotic twins). In that case, we can consider a set of relationships $\mathcal{A}$ and test R_o against $\mathcal{A} \setminus R_o$. The test statistic could be the maximized log-likelihood ratio ($MLLR$),

$$MLLR = \log(\hat{L}_{\mathcal{A}}) - \log(L_{R_o}),$$

$$\hat{L}_{\mathcal{A}} = \max_{R \in \mathcal{A} \setminus R_o} L_R.$$

The statistic $MLLR$ has a rather skewed distribution, and it requires simulation to assess significance. The empirical p-value of the $MLLR$ test can be calculated by simulating genotype data under the null relationship.

The test statistic $MLLR$ depends on the particular set $\mathcal{A}$ considered. In order for the test to have high power, it is desirable that the true relationship for the pair to be tested is included in $\mathcal{A}$. However, for each relationship $R \in \mathcal{A}$, we need to augment a Markov process and apply a hidden Markov model to calculate the likelihood. For data in which thousands of pairs are to be examined, considering all possible pairwise relationships may not be computationally feasible. To implement the $MLLR$ test for outbred pedigrees, We take $\mathcal{A}$ to include 11 relationships: MZ-twin, parent-offspring, full-sib, half-sib-plus-first-cousin (Figure 2.6), half-sib, grandparent-grandchild, avuncular (Figure 2.1), first-cousin (Figure 2.3), half-avuncular (Figure 2.4), half-first-cousin (Figure 2.5) and unrelated. Based on our experience with data analysis, the majority of pairs in a typical pedigree collected for linkage studies fit into these 11 relationships (Chapter 4). Other relationships may be approximated by those in $\mathcal{A}$. For instance, the first-cousin-once-removed relationship is well-approximated by the half-first-cousin relationship. Note that one may detect a misspecified relationship using the $MLLR$ test, even when the true relationship is not in $\mathcal{A}$.

3.1.2 The IBS test

Ehm and Wagner (1998) propose a simple test statistic, which we call IBS', based the number of alleles shared IBS: $IBS' = \Sigma_m X_m/2$, where X_m is the number of alleles shared IBS by a pair at marker m. IBS' is approximately normally distributed if it is applied to a large number of markers. In the case of a full-sib pair, Ehm and Wagner (1998) describe the calculation of the null mean and variance of IBS' by conditioning on the mating type of the parents. For more general outbred pairs, we consider the statistic

$$IBS = \frac{1}{n}\sum_{m=1}^{n} X_m,$$

where n is the total number of markers, and X_m is as defined above. We verify by simulations that the normal approximation works well for assessing significance of IBS, when genotype data are from a genome-screen. To calculate the null mean $E_{R_o}[IBS]$ and the null variance $Var_{R_o}(IBS)$, we develop a method that is applicable to non-sib pairs. Let $(f_1, f_2, \ldots, f_l)$ be the frequency distribution of l alleles at marker m. In addition to the null IBD probabilities $p_i = P_{R_o}(D_m = i)$, $i \in \{0,1,2\}$, the calculation of $E_{R_o}[IBS]$ and $Var_{R_o}(IBS)$ requires the following probabilities, valid for any outbred pair:

$$\begin{aligned}
&P(X_m = 2|D_m = 2) = 1,\\
&P(X_m = 2|D_m = 1) = \textstyle\sum_i f_i^2,\\
&P(X_m = 1|D_m = 1) = 1 - \textstyle\sum_i f_i^2,\\
&P(X_m = 2|D_m = 0) = \textstyle\sum_i f_i^4 + 2\sum_i \sum_{j\neq i} f_i^2 f_j^2,\\
&P(X_m = 1|D_m = 0) = 4\textstyle\sum_i \sum_{j\neq i} f_i^3 f_j + 4\sum_i \sum_{j\neq i} \sum_{k\neq i,j} f_i^2 f_j f_k,\\
&P(X_{m_1}, X_{m_2}|D_{m_1} = i1, D_{m_2} = i2) = P(X_{m_1}|D_{m_1} = i1)P(X_{m_2}|D_{m_2} = i2),
\end{aligned}$$

where $m_1 \neq m_2$, $i1, i2 \in \{0,1,2\}$. The null two-locus IBD transition probabilities $P_{R_o}(D_{m_2} = i2|D_{m_1} = i1)$, $i1, i2 \in \{0,1,2\}$ are also needed. The calculation of $P_R(D_{m_2}|D_{m_1})$ for general relationships is discussed in Section 2.2.5. In particular, the probabilities for full-sib, half-sib, parent-offspring, grandparent-grandchild, avuncular and first-cousin pairs are given in Table 1 of Bishop and Williamson (1990). We have derived the probabilities for for half-avuncular, half-first-cousin and half-sib-plus-first-cousin pairs. The results are shown, respectively, in Table 2.10, Table 2.14 and Table 2.15. (Note that $P(D_{m_2} = 2|D_{m_1} = 2) \equiv 1$ for MZ-twin pairs, $P(D_{m_2} = 1|D_{m_1} = 1) \equiv 1$ for parent-offspring pairs, and $P(D_{m_2} = 0|D_{m_1} = 0) \equiv 1$ for unrelated pairs.) Given the above probabilities, we have

$$E_{R_o}[IBS] = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=0}^{2}\sum_{j=0}^{2} jp_iP(X_m = j|D_m = i),$$

$$Var_{R_o}(IBS) = \left[\frac{1}{n^2}\sum_{m_1=1}^{n}\sum_{m_2=1}^{n}\sum_{i1=0}^{2}\sum_{i2=0}^{2}\sum_{j1=0}^{2}\sum_{j2=0}^{2} j1j2p_{i1} \times P_{R_o}(D_{m_2} = i2|D_{m_1} = i1)\times\right.$$

$$\left. P(X_{m_1} = j1|D_{m_1} = i1) \times (P(X_{m_2} = j2|D_{m_2} = i2))^{I(m_1 \neq m_2)}\right] - E_{R_o}[IBS]^2,$$

where $I\{m_1 \neq m_2\}$ is the indicator of the event $\{m_1 \neq m_2\}$.

3.1.3 The $EIBD$ and $AIBS$ tests

While the $MLLR$ test has high power, its drawbacks include the need to construct Markov processes for all relationships considered in $\mathcal{A}$ and implement a separate hidden Markov calculation for each, and the need for simulation to assess significance. Thus, the $MLLR$ test may be very cumbersome to use as a diagnostic tool. The IBS test is much simpler computationally, but it loses power by ignoring chance sharing of common alleles. Consider a marker with 2 alleles, $a1$ with frequency $f_{a1} = 1 - \epsilon$ and $a2$ with frequency $f_{a2} = \epsilon$ for ϵ small, and consider two possible pairs of genotypes ($a1$ $a1$ $a1$ $a1$) and ($a2$ $a2$ $a2$ $a2$) for a pair of individuals. (In each case, the first two alleles are genotype for one individual and the second two alleles are genotype for the other individual.) Because $a1$ is a very common allele and $a2$ is an extremely rare allele, it is likely that two unrelated individuals have genotypes $(a1\ a1\ a1\ a1)$ and unlikely that they have genotypes $(a2\ a2\ a2\ a2)$. Thus, $(a2\ a2\ a2\ a2)$ suggests a closer relationship than $(a1\ a1\ a1\ a1)$ does. However, test statistic $IBS = 2$ for both cases. Not surprisingly, the power of the test based on IBS is low.

To compromise between $MLLR$ and IBS, we propose two new statistics designed to retain the computational simplicity of IBS but to increase power by taking into account chance sharing. The first test statistic, denoted $EIBD$, is the average of the conditional expected number of alleles shared IBD by a pair of individuals at each marker, conditional on the observed genotype data at that marker and the null relationship for the pair,

$$EIBD = \frac{1}{n} \sum_{m=1}^{n} E_{R_o}[D_m | G_m],$$

where D_m is the number of alleles shared IBD at marker m, G_m is the genotype data for the pair at marker m, n is the total number of markers, and the subscript R_o indicates that the expectation is calculated under the null relationship. In the case of an outbred pair,

$$E_{R_o}[D_m | G_m] = \frac{P(G_m | D_m = 1)p_1 + 2P(G_m | D_m = 2)p_2}{\sum_{i=0,1,2} P(G_m | D_m = i)p_i},$$

where $p_i = P_{R_o}(D_m = i)$, $i \in \{0, 1, 2\}$ are the null IBD probabilities for the pair. The probabilities $P(G_m | D_m = i)$, $i \in \{0, 1, 2\}$ are given in Thompson (1975). Now consider the previous example in which there are two different pairs of genotypes, $(a1\ a1\ a1\ a1)$ and $(a2\ a2\ a2\ a2)$ with $f_{a1} = 1 - \epsilon$ and $f_{a2} = \epsilon$ for ϵ small. Note that $P((ai\ ai\ ai\ ai) | D = 0) = f_{ai}^4$, $P((ai\ ai\ ai\ ai) | D = 1) = f_{ai}^3$, and $P((ai\ ai\ ai\ ai) | D = 2) = f_{ai}^2$, $i \in \{1, 2\}$. Thus, $E_{R_o}[D_m | (a1\ a1\ a1\ a1)]$ approaches its null expected value of $p_1 + 2p_2$ as $\epsilon \to 0$, which is appropriate because observation $(a1\ a1\ a1\ a1)$ provides almost no information on the IBD sharing of the alleles. In contrast, $E_{R_o}[D_m|(a2\ a2\ a2\ a2)] \to 2$ as $\epsilon \to 0$, because observation $(a2\ a2\ a2\ a2)$ provides almost complete information on the IBD sharing of the alleles. Our simulations indicate that the normal distribution gives a close approximation to the sampling distribution of $EIBD$ when applied to genome-screen data. Thus, to assess significance, one needs only to compute the null mean and variance of the statistic. The null mean $E_{R_o}[EIBD]$ has a simple close form,

$$E_{R_o}[EIBD] = E_{R_o}[E_{R_o}(D_m|G_m)] = E_{R_o}[D_m] = p_1 + 2p_2 = 4\Phi,$$

where Φ is the kinship coefficient defined in Section 1.3. The calculation of the null variance $Var_{R_o}[EIBD]$ is very similar to that for IBS. We can think of $E_{R_o}(D_m|G_m)$ as a function of G_m. Then we need the probabilities $P_{R_o}(D_{m_2} = i2|D_{m_1} = i1)$, $i1, i2 \in \{0, 1, 2\}$ as in the IBS case and the probabilities $P(G_m|D_m = i)$, $i \in \{0, 1, 2\}$.

One possible drawback of $EIBD$ is that, if the null relationship has $p_2 = 0$, then $E_{R_o}(D_m|G_m)$ is restricted to lie between 0 and 1. This may give less than optimal power if the true relationship has moderate p_2. To avoid this problem, we also propose an adjusted IBS statistic ($AIBS$), which is an average of Y_m over all markers m,

$$AIBS = \frac{1}{n}\sum_{m=1}^{n} Y_m, \text{ where}$$

$$\begin{aligned}
&Y_m = 0, \quad \text{if } G_m = (i\ \ j\ \ k\ \ l),\ i,j \neq k,\ i,j \neq l,\\
&Y_m = \tfrac{\Phi_{R_o}}{\Phi_{R_o}+(1-\Phi_{R_o})f_i}, \quad \text{if } G_m = (i\ \ j\ \ i\ \ k),\ j \neq k,\\
&Y_m = \tfrac{\Phi_{R_o}}{\Phi_{R_o}+(1-\Phi_{R_o})f_i} + \tfrac{\Phi_{R_o}}{\Phi_{R_o}+(1-\Phi_{R_o})f_j}, \quad \text{if } G_m = (i\ \ j\ \ i\ \ j),
\end{aligned}$$

where f_i is the frequency of allele i, and Φ_{R_o} is the kinship coefficient for the pair under the null relationship. If an allele are drawn at random from each individual's genotype at a given locus, the quantity $\Phi_{R_o}/(\Phi_{R_o} + (1 - \Phi_{R_o})f_i)$ would represent the probability that the two alleles are shared IBD given that they are shared IBS for allele i, conditional on the null relationship. Our simulations show that that the normal approximation is quite satisfactory for assessing significance of $AIBS$. The calculations of the null mean and variance of $AIBS$ are very similar to those of IBS and $EIBD$, because Y_m can be thought of as a function of G_m.

3.2 Power and robustness studies

We perform simulations to compare power of the $MLLR$, $EIBD$, $AIBS$ and IBS tests for detection of misspecified pairwise relationships. We also include in the comparison the LLR test using the correct alternative relationship, which sets a benchmark of close to optimal power that is not realistically achievable in practice when the correct alternative relationship is unknown. (Power of LLR used here is slightly suboptimal because of the presence of interference, but we expect the effect to be almost negligible. See Section 2.2.4 for the extension of the likelihood calculation to the case of interference.) In our initial simulation, we consider the following eleven relationships: MZ-twin, parent-offspring, full-sib, half-sib-plus-first-cousin, half-sib, grandparent-grandchild, avuncular, first-cousin, half-avuncular, half-first-cousin and unrelated. We simulate marker data from an autosomal genome-screen for which we vary the allele frequencies and marker resolution. Our simulate scenarios include panels of microsatellite markers equally spaced at recombination fractions of 0.07, 0.15 and 0.25, with sex-averaged chromosome length taken from

Broman *et al.* (1998), and with all markers having the allele frequency distribution, $(0.40, 0.20, 0.20, 0.05, 0.05, 0.05, 0.05)$. We also simulate single nucleotide polymorphisms (SNPs) equally spaced at recombination fractions of 0.01 and 0.07, with the same allele frequency distribution, $(0.7, 0.3)$. The allele frequencies for these simulated SNP and microsatellite panels are chosen so that the markers would be somewhat less informative than the ideal, but within the range of what might be typical. Our results show that the conclusions about power comparisons across the statistics depend very little on the assumptions about the allele frequency distributions. Our final simulated marker panel is based on the markers actually typed in the GAW 11 COGA data (Section 4.1). This panel is more realistic because marker spacings are unequal with an average intermarker recombination fraction of 0.13, allele frequency distributions differ across markers, and some marker data are missing. We consider the power of the hypothesis tests at significance levels of 0.05, 0.01, 0.005 and 0.001. The significance level of 0.05 or 0.01 would be appropriate for a single hypothesis test. However, in practice, we often need to check for pedigree errors in a moderate number of pedigrees with a large number of pairs. This creates a problem of multiple comparisons. To reduce the large number of false positives that would be expected, a lower significance level is recommended. In that case, the significance level of 0.005 or 0.001 could be used to "flag" pairs that are problematic (discussed further in Chapter 4).

All simulations are performed using the χ^2 interference model for crossovers with interference with parameter m=4 for humans, corresponding to a gamma shape parameter of 5, as suggested by the results of Lin and Speed (1996). Although it is convenient to assume no interference in the development and implementation of the testing methods, the actual data do contain interference. Thus, in order to give as close an indication as possible of the performance of the methods on real data, we simulate the data with interference. The number of replications used to assess the power is 1 million. The five testing methods are analyzed on the same 1 million data sets in each case, minimizing any effects of sampling variability. For the $EIBD$, $AIBS$ and IBS tests, the normal approximation is adequate to asses significance. This does not hold for the $MLLR$ test, for which simulation is required to assess the empirical p-value. In that case, 10^5 realizations are generated. For the $EIBD$, $AIBS$ and IBS tests, we compare the p-values calculated from the normal approximation to those calculated from the empirical null distributions and found them to be very close (results not shown). We also found that the distributions of these test statistics are approximately normal even when the null relationship is not the true relationship.

Figure 3.1 and Tables 3.1, 3.2 and 3.3 give some of the results of the power studies. Figure 3.1 to show the power of the tests (at significance level 0.001) for one case, in which the genotype data are simulated from the first-cousin relationship and the null hypothesis is assumed to be the half-sib relationship, for three idealized maps and for the COGA map. The idealized maps consist of microsatellite markers evenly spaced at recombination fractions of 0.07, 0.15 and 0.25, with all markers having the allele frequency distribution, $(0.40, 0.20, 0.20, 0.05, 0.05, 0.05, 0.05)$. From Figure 3.1, it is clear that the LLR and $MLLR$ tests have the highest power, followed by the $EIBD$ and $AIBS$ tests, with the IBS test having lower power than the others. Note that the power achieved by the $MLLR$ test is close to the optimal power set by the LLR test. This ordering of power is true for most of the cases simulated. We point out that although the IBS test has the lowest power among the tests considered, it can

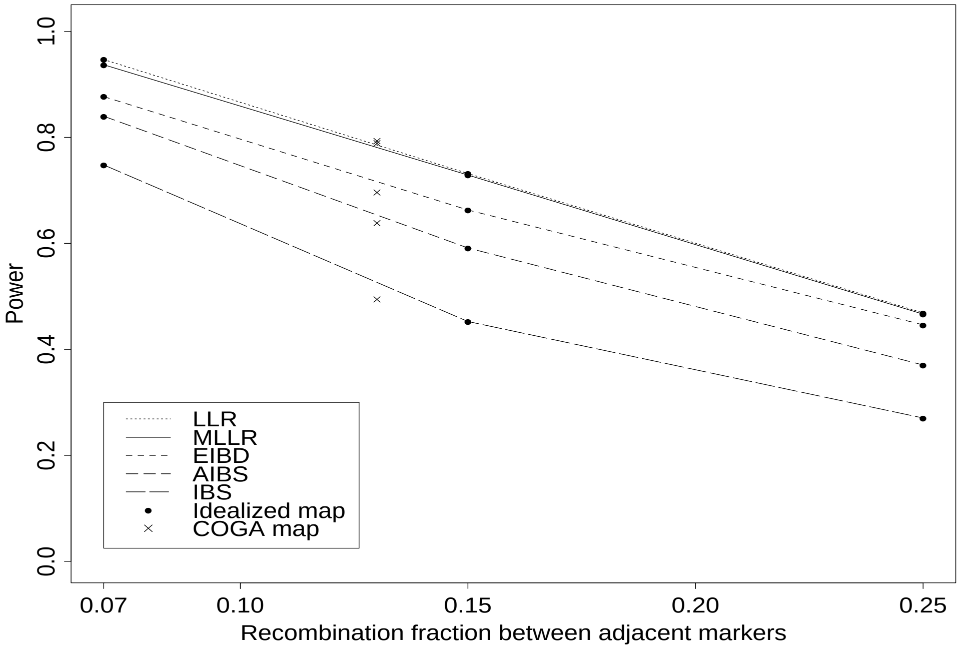

Figure 3.1: Power versus genome-screen resolution

be useful for quickly identifying MZ twins or duplicated samples. In those two cases, the value of observed statistic IBS will be either exactly two, or just slightly below two if genotyping errors occurred. The results are similar to those in Figure 3.1 if the null of the first-cousin relationship is tested against the alternative of the half-sib or avuncular relationship (see Table 3.1), or the null of the avuncular relationship is tested against the alternative of the first-cousin relationship (results not shown). Table 3.1 also gives power for testing the null of the first-cousin relationship against the alternative of the grandparent-grandchild relationship for significance level 0.001. The ordering of the five statistics in terms of their power is the same as above.

Marker type	Test	Alternative (true) relationship		
(θ)	statistic	half-sib	grandparent-grandchild	avuncular
SNP	LLR	.99	.99	.98
(.01)	$MLLR$	.98	.99	.98
	$EIBD$	.95	.91	.95
	$AIBS$	.89	.85	.89
	IBS	.88	.85	.89
microsatellite	LLR	.96	.99	.95
(.07)	$MLLR$	.96	.98	.95
	$EIBD$	.91	.87	.92
	$AIBS$	.86	.82	.87
	IBS	.79	.76	.79
COGA map	LLR	.82	.90	.81
(avg. .13)	$MLLR$	.82	.88	.80
	$EIBD$	.75	.73	.76
	$AIBS$	.65	.63	.65
	IBS	.54	.54	.54
microsatellite	LLR	.77	.86	.75
(.15)	$MLLR$	.76	.84	.74
	$EIBD$	.71	.69	.72
	$AIBS$	.59	.58	.59
	IBS	.45	.46	.45
microsatellite	LLR	.49	.57	.48
(.25)	$MLLR$	.48	.56	.47
	$EIBD$	.47	.47	.47
	$AIBS$	.35	.35	.34
	IBS	.23	.24	.23
SNP	LLR	.40	.48	.39
(.07)	$MLLR$	.39	.47	.37
	$EIBD$	.36	.37	.36
	$AIBS$	.19	.20	.18
	IBS	.18	.19	.18

Table 3.1: Power of tests based on LLR, $MLLR$, $EIBD$, $AIBS$ and IBS

The null hypothesis of the tests is the full-cousin relationship. The number of replications used to assess the power is 1 million. The number of replications used to assess the empirical p-values of LLR and $MLLR$ is 10^5. Significance level is 0.001. Microsatellite markers have allele frequency distribution $(.40, .20, .20, .05, .05, .05, .05)$, and SNPs have allele frequency distribution $(.7, .3)$, and markers are equally spaced with given recombination fraction θ between adjacent pairs. GAW 11 COGA map has an average marker spacing of about 0.13.

The half-sib, avuncular and grandparent-grandchild relationships are similar in that they have the same IBD probabilities, $\mathbf{p} = (p_0, p_1, p_2) = (1/2, 1/2, 0)$. However, the transition rate of the grandparent-grandchild IBD process is only 1/2 that of half-sibs and 2/5 that of avuncular. (In each case, we define the transition rate to be the rate of transition to IBD value $1 - i$ conditional on current IBD value i for the stationary IBD process.) The LLR and $MLLR$ tests are the only ones among the five tests considered that take into account the information in the data on the transition rate of the process. For each of the statistics $EIBD$, $AIBS$ and IBS, its mean value does not vary among the three relationships, although its variance does vary. Thus, the tests based on these statistics have almost no power to distinguish among these three relationships. The LLR and $MLLR$ tests have some power to distinguish among them based mainly on the different transition rates of the IBD processes. However, as shown in Table 3.2, Even the most powerful LLR test has very low power in these cases. Note that the power is higher when one in the pair of relationships (null or alternative) is grandparent-grandchild than when neither relationship is grandparent-grandchild. This is explained by the fact that the transition rate for the grandparent-grandchild IBD process is very different from those for the avuncular and half-sib IBD processes.

We find that the full-sib relationship is relatively easy to distinguish from the other relationships, either as a null or as an alternative. Tables 3.3 show power when the full-sib relationship is either the null or the alternative, for significance level 0.001, for microsatellites with recombination fraction between adjacent markers of 0.25 and for SNP's with recombination fraction between adjacent markers of 0.07. For lower significance levels or increased marker density, power is nearly perfect for

Marker type	Alternative (true) relationship	Null (false) relationship		
(θ)		half-sib	grandparent-grandchild	avuncular
SNP	half-sib	NA	.47	.01
(.01)	grandparent-grandchild	.40	NA	.73
	avuncular	.02	.78	NA
microsatellite	half-sib	NA	.23	.01
(.07)	grandparent-grandchild	.20	NA	.41
	avuncular	.01	.47	NA
COGA map	half-sib	NA	.06	.03
(avg. .13)	grandparent-grandchild	.05	NA	.10
	avuncular	.00	.13	NA
microsatellite	half-sib	NA	.04	.00
(.15)	grandparent-grandchild	.04	NA	.07
	avuncular	.00	.09	NA
microsatellite	half-sib	NA	.01	.00
(.25)	grandparent-grandchild	.01	NA	.02
	avuncular	.00	.02	NA
SNP	half-sib	NA	.01	.00
(.07)	grandparent-grandchild	.01	NA	.01
	avuncular	.00	.02	NA

Table 3.2: Power of LLR to distinguish half-sib, grandparent-grandchild and avuncular relationships NA means not applicable. Note: for details, see legend of Table 3.1.

all methods when the full-sib relationship is either the null or the alternative, so the results are not shown. Note that when the alternative relationship is full-sib while the null relationship has $p_2 = 0$, $EIBD$ performs worse than the other statistics. This is because the conditional expected number of alleles shared IBD can never exceed 1 in that case. Even so, the power of $EIBD$ in this case is very high, at least 88% for all cases simulated.

The type of map used, SNP map, microsatellite map or the COGA map, does not have a substantial impact on the power comparisons among the statistics. The power results using SNPs with allele frequency distribution $(0.7, 0.3)$ are similar to those using microsatellites at a lower density. Using SNPs at recombination fraction 0.01 generally gives a test with slightly more power than that using microsatellites at recombination fraction 0.07, while using SNPs at recombination fraction 0.07 generally gives a test with slightly less power than that using microsatellites at recombination fraction 0.25. When SNPs with allele frequency distribution $(0.5, 0.5)$ are used, power is slightly higher, but the increase is fairly small (results not shown). The results for the COGA map, which has average intermarker recombination fraction of 0.13 and different allele frequency distributions across markers, are quite similar to those for the idealized microsatellite map with intermarker recombination fraction of 0.15.

In most of our simulations, we assume that allele frequencies and marker map positions are known, whereas in practice, one would generally need to estimate these from available genotype data. We perform simulations to study how robust each test is to the misspecification of allele frequencies or genetic distances between markers. Our preliminary results suggest that the $MLLR$ test is more robust to the misspecified allele frequencies than are the $EIBD$, $AIBS$ and IBS tests, but the latter three

Marker type	Test	Relationship			
(θ)	statistic	half-sib	grandparent-grandchild	avuncular	first-cousin
			being null (false)	relationship	
microsatellite (.25)	LLR	1.00	1.00	1.00	1.00
	$MLLR$	0.99	0.99	1.00	1.00
	$EIBD$	0.92	0.88	0.93	1.00
	$AIBS$	1.00	0.99	1.00	1.00
	IBS	1.00	0.99	1.00	1.00
SNP (.07)	LLR	1.00	1.00	1.00	1.00
	$MLLR$	1.00	1.00	1.00	1.00
	$EIBD$	0.95	0.92	0.96	1.00
	$AIBS$	0.99	0.99	1.00	1.00
	IBS	0.99	0.99	1.00	1.00
			being alternative (true)	relationship	
microsatellite (.25)	LLR	1.00	1.00	1.00	1.00
	$MLLR$	1.00	1.00	1.00	1.00
	$EIBD$	1.00	1.00	1.00	1.00
	$AIBS$	1.00	1.00	1.00	1.00
	IBS	1.00	0.99	1.00	1.00
SNP (.07)	LLR	1.00	1.00	1.00	1.00
	$MLLR$	1.00	1.00	1.00	1.00
	$EIBD$	1.00	1.00	1.00	1.00
	$AIBS$	1.00	0.99	1.00	1.00
	IBS	0.99	0.99	0.99	1.00

Table 3.3: Power of the tests for case of the full-sib relationship

Top part of the table: power of tests based on LLR, $MLLR$, $EIBD$, $AIBS$ and IBS against alternative of the full-sib relationship when null is the half-sib, grandparent-grandchild, avuncular, or first-cousin relationships. Bottom part of the table: power the tests against alternative of the half-sib, grandparent-grandchild, avuncular, or first-cousin relationship when null is the full-sib relationship. Significance level is 0.001. Note: for details on marker type, see legend of Table 3.1.

tests are more robust to the misspecified marker map positions than is the $MLLR$ test.

3.3 Relationship estimation

If the null relationship for a pair is rejected in a hypothesis test, it is of interest to know what relationships are suggested by the observed genotype data for the pair. One strategy is to create augmented Markov chains for a variety of relationships, calculate the likelihood under each, and consider the relationships with the largest likelihoods. However, one would need to first guess the correct relationship in order to construct an augmented Markov for it. Furthermore, the need to specify and implement a separate Markov process for each relationship considered would involve a substantial investment of computational time. We propose a simpler strategy, involving estimation of the probability distribution of the IBD states, i.e. estimate $\mathbf{p} = (p_0, p_1, p_2)$. We estimate $\mathbf{p}$ by maximizing

$$\sum_{m=1}^{n} \log(L(G_m; \mathbf{p})),$$

where $L(G_m; \mathbf{p})$ is the likelihood of the genotype data at marker m in terms of $\mathbf{p}$,

$$L(G_m; \mathbf{p}) = p_0 P(G_m | D_m = 0) + p_1 P(G_m | D_m = 1) + p_2 P(G_m | D_m = 2).$$

If the markers are unlinked, this estimate of $\mathbf{p}$ would be the maximum likelihood estimate (MLE) derived in Thompson (1975). Here we apply this procedure to linked markers. The quantity $\sum_{m=1}^{n} \log(L(G_m; \mathbf{p}))$ can be quickly maximized by an application of the EM algorithm. Where the current estimate of $\mathbf{p}$ is $\mathbf{p}^{(k)} = (p_0^{(k)}, p_1^{(k)}, p_2^{(k)})$,

$p_0^{(k)} + p_1^{(k)} + p_2^{(k)} = 1$, we obtain the updated estimate by the following formula:

$$p_i^{(k+1)} = \frac{p_i^{(k)}}{n} \sum_{m=1}^{n} P(G_m | D_m = i)/L(G_m; \mathbf{p}^{(k)}).$$

As shown in Thompson (1986), under the assumption of no inbreeding, the constraint $p_1^2 \geq 4p_0p_2$ must be satisfied. The quantity $\sum_{m=1}^{n} \log(L(G_m; \mathbf{p}))$ could be maximized subject to this constraint by first finding the maximizing $\mathbf{p}$ in the unconstrained case. If the constraint is violated, then the condition $p_1^2 = 4p_0p_2$ is imposed and a one-dimensional search algorithm is used. However, because the true relationship is not known, one may want to use the unconstrained estimate to allow for the possibility of inbreeding. We perform simulations to investigate the properties of this estimator when applied to genome-screen data. Some of the results are given in Table 3.4. For the relationships with $p_2 = 0$, there is a slight bias in the estimates, amounting to no more than about 5%. This bias is expected because one can estimate p_2 only at or above its actual value, never below, in those cases. For microsatellite markers at recombination fraction 0.07, the bias is quite small. The standard deviations of the estimates tend to be rather large at the marker resolutions considered. Thus, the procedure may give only a rough idea of the true relationship. However, in some cases, the estimate of $\mathbf{p} = (p_0, p_1, p_2)$ can be very useful for proposing some likely relationships for a pair. The proposed relationships can then be tested for fit to data. For example, consider a case where the estimate of $\mathbf{p}$ for a putative full-sib pair is $(.463, .512, .025)$. The true value of $\mathbf{p}$ is $(.25, .5, .25)$ for a full-sib pair and $(.5, .5, 0)$ for a half-sib, avuncular or grandparent-grandchild pair, so it might be reasonable to test whether these relationships are compatible with the observed genotype data for this putative full-sib pair. Note that, by inverting the hypothesis test, a confidence set of relationships compatible with the data for the given pair can be constructed by including in the set all relationships not rejected.

3.4 Implementation

We have implemented all of the methods described in Sections 3.1 and 3.3 as software consisted of a pair of C programs, PREST and ALTERTEST. This software was developed in the hope that it will be useful for researchers who want to detect pedigree errors quickly and effectively before carrying out their genetic mapping studies. The documentation of the software and the two programs are all freely available on the web at `http://galton.uchicago.edu/∼mcpeek/software/prest`. We describe the main functions of the two programs in Appendix D. Both programs have been successfully applied to several data sets including the GAW 11 COGA data set (Section 4.1), and the GAW 12 CSGA data set and GER data set (Sun, Abney and McPeek, in press). The software is applicable to most of the data collected for linkage studies, but it is not directly applicable to extreme data such as the Hutterites, in which there are no pairs which fit into the relationships that have been implemented. In that case, we used a graphical method based on the extended test statistic $EIBD$ (see Appendix C) to detect errors in the Hutterite pedigree (Section 4.2).

Marker type	Relationship	Mean (standard deviation) of estimates		
(θ)		p_0	p_1	p_2
SNP	full-sib	.251 (.045)	.499 (.050)	.250 (.042)
(.01)	half-sib	.504 (.057)	.488 (.060)	.008 (.011)
	grandparent-grandchild	.505 (.068)	.488 (.070)	.008 (.011)
	avuncular	.505 (.055)	.488 (.057)	.007 (.011)
	first-cousin	.755 (.051)	.238 (.054)	.007 (.010)
microsatellite	full-sib	.249 (.047)	.500 (.052)	.251 (.044)
(.07)	half-sib	.502 (.060)	.490 (.061)	.007 (.011)
	grandparent-grandchild	.502 (.070)	.491 (.071)	.007 (.011)
	avuncular	.503 (.057)	.490 (.058)	.007 (.011)
	first-cousin	.752 (.054)	.242 (.056)	.006 (.009)
COGA map	full-sib	.249 (.053)	.501 (.062)	.250 (.048)
(avg. .13)	half-sib	.503 (.068)	.487 (.071)	.009 (.014)
	grandparent-grandchild	.504 (.077)	.487 (.079)	.009 (.014)
	avuncular	.503 (.066)	.487 (.069)	.009 (.014)
	first-cousin	.754 (.065)	.238 (.067)	.008 (.012)
microsatellite	full-sib	.250 (.055)	.500 (.064)	.250 (.048)
(.15)	half-sib	.503 (.070)	.486 (.073)	.010 (.016)
	grandparent-grandchild	.503 (.079)	.486 (.081)	.010 (.016)
	avuncular	.504 (.068)	.485 (.071)	.011 (.016)
	first-cousin	.756 (.069)	.235 (.071)	.009 (.014)
microsatellite	full-sib	.250 (.065)	.500 (.080)	.251 (.055)
(.25)	half-sib	.503 (.084)	.483 (.088)	.013 (.021)
	grandparent-grandchild	.504 (.090)	.482 (.094)	.014 (.021)
	avuncular	.505 (.083)	.481 (.087)	.014 (.021)
	first-cousin	.758 (.082)	.231 (.086)	.012 (.018)
SNP	full-sib	.250 (.074)	.500 (.097)	.250 (.059)
(.07)	half-sib	.510 (.093)	.470 (.102)	.019 (.029)
	grandparent-grandchild	.513 (.101)	.468 (.110)	.020 (.029)
	avuncular	.513 (.093)	.468 (.102)	.019 (.028)
	first-cousin	.766 (.094)	.215 (.103)	.019 (.027)

Table 3.4: Relationship estimation results

Each mean of estimated IBD sharing probability is based on 10^4 simulated realizations, with the standard deviation of the estimate in parentheses. p_i is the probability of i alleles shared IBD, $i = 0, 1, 2$. Note: for details on marker type, see legend of Table 3.1.

3.5 Appendix C

In principle, all the methods described in Sections 3.1 and 3.3 can be extended to the case of inbred relationships. The likelihood calculation for an outbred relative pair described in Section 2.2.2 would allow construction of a likelihood ratio test for outbred relationships. However, the likelihood approach is difficult for inbred pedigrees, and it becomes computationally infeasible for the Hutterite pedigree.

In order to extend the definition of $EIBD$ to inbreds, there is more than one reasonable approach. One could define the of number of alleles shared IBD for an inbred relative pair by defining states $(S_1, ..., S_9)$ (see Figure 2.2) to have $(4, 0, 2, 0, 2, 0, 2, 1, 0)$ alleles shared IBD. Alternatively, one might prefer to define states $(S_1, ..., S_9)$ to have $(2, 0, 1, 0, 1, 0, 2, 1, 0)$ alleles shared IBD. Then, one could define $EIBD$ to be

$$EIBD = \frac{1}{n} \sum_{m=1}^{n} E_{R_o}[D_m|G_m], \text{ where}$$

$$E_{R_o}[D_m|G_m] = \frac{\sum_{i=1}^{9} S_i P(G_m|D_m = S_i)\Delta_i}{\sum_{i=1}^{9} P(G_m|D_m = S_i)\Delta_i},$$

where $(\Delta_1, ..., \Delta_9)$ are the null probabilities of the nine identity states $(S_1, ..., S_9)$ for the pair. The probabilities $P(G_m|D_m = S_i)$, $S_i \in \{S_1, ..., S_9\}$ are given in Jacquard (1970). In application, we choose to assign states $(S_1, ..., S_9)$ to have $(4, 0, 2, 0, 2, 0, 2, 1, 0)$ alleles shared IBD. This definition ensures that the equation $E_{R_o}[EIBD] = 4\Phi$ holds as in the outbred case,

$$E_{R_o}[EIBD] = 4\Delta_1 + 2(\Delta_3 + \Delta_5 + \Delta_7) + \Delta_8 = 4\Phi.$$

To extend the definition of IBS to inbreds, one could think of 9 IBS states analogous to the 9 IBD states. Again, to apply the IBS score statistic to inbreds, one could define the number of alleles shared IBS for inbred relative pairs by defining IBS states $(S_1, ..., S_9)$ to have $(4, 0, 2, 0, 2, 0, 2, 1, 0)$ alleles shared IBS. However, if the level of inbreeding is low, one may have more power by defining IBS states $(S_1, ..., S_9)$ to have $(2, 0, 1, 0, 1, 0, 2, 1, 0)$ alleles shared IBS. The definition of $AIBS$ statistic for an inbred relative pair is the same as that for the outbred case, with kinship coefficient Φ replaced by $\Delta_1 + (\Delta_3 + \Delta_5 + \Delta_7)/2 + \Delta_8/4$. We note that, for a pair of inbred individuals, the calculations of the null variances of test statistics IBS, $EIBD$ and $AIBS$ are generally computationally intensive, and they becomes computationally infeasible for pairs in the Hutterite pedigree.

To perform pairwise relationship estimation for inbred relative pairs, in principle, we could estimate $\Delta = (\Delta_1, ..., \Delta_9)$. However, the level of inbreeding is generally low in human pedigrees, and genotype data may not provide enough information to distinguish, for example between Δ_2 and Δ_9. Thus, for inbred pairs, we estimate combined parameters (also denoted as p_i for simplicity): $p_0 = \Delta_2 + \Delta_4 + \Delta_6 + \Delta_9$, $p_1 = \Delta_8$, and $p_2 = \Delta_1 + \Delta_3 + \Delta_5 + \Delta_7$. p_0, p_1 and p_2 can be estimated using the same method described in Section 3.3.

3.6 Appendix D

We have implemented all of the methods described in Sections 3.1 and 3.3 as software consisting of a pair of C programs, PREST and ALTERTEST. The software is freely available at `http://galton.uchicago.edu/~mcpeek/software/prest`.

The PREST program detects possible misspecified relationships in general outbred pedigrees. ("PREST" stands for Pedigree Relationship Statistical Test). First, based on the information provided by the given pedigree(s), PREST identifies all instances of parent-offspring, full-sib, half-sib-plus-first-cousin, half-sib, grandparent-grandchild, avuncular, first-cousin, half-avuncular, half-first-cousin and unrelated pairs within each pedigree. (The standard input format for pedigree data usually does not specify MZ twins.) Then, PREST gives the user two options. As discussed before, although the $MLLR$ test has good power, one of its drawbacks is the need for simulation to assess significance. In data such as the GAW 11 COGA data (Section 4.1), there may be thousands of pairs to be tested, and the $MLLR$ test in that case can be time-consuming and insufficient. For instance, for each pair in the COGA data, the $MLLR$ test using 10^5 simulated realizations takes approximately 5 minutes on a Sun Ultra-2. In contrast, for each pair, the $EIBD$, $AIBS$ and IBS tests using the normal approximation take less than 80 milliseconds on the same machine. For a more practical approach, one could use the $EIBD$, $AIBS$ and IBS tests as preliminary screening tools to determine a subset of pairs, $\mathcal{R}$, for which it is worthwhile to perform the more time-consuming but powerful $MLLR$ test, or do multiple stages of $MLLR$ simulations, with more replicates for more significant results. We have implemented the first approach as a two-stage screening procedure, which can be summarized as follows: In stage one, a putative parent-offspring pair is placed in $\mathcal{R}$ if $\hat{p}_1 < 0.9$. A putatively unrelated pair is placed in $\mathcal{R}$ if the p-value of the IBS test is < 0.2. (The $EIBD$ and $AIBS$ tests are not applicable to an unrelated pair.) Any other type of relative pair is placed in $\mathcal{R}$ if at least one of the following holds: (i) the p-value of the $EIBD$ test is < 0.2; (ii) the p-value of the $AIBS$ test is < 0.2; (iii)

$\hat{p}_1 > 0.75$. In stage two, the $MLLR$ test is applied to each pair in $\mathcal{R}$. Thus, if option one is chosen, P REST will perform stage one of the two-stage screening procedure as described above. (In stage one, for each pair identified, the $EIBD$, $AIBS$ and IBS tests are performed, $\mathbf{p} = (p_0, p_1, p_2)$ is estimated, and a selection of pairs based on the combined testing and estimation results is applied.) If option two is chosen, PREST will perform both stage one and stage two. (In stage two, for each pair that passed the screening criteria in stage one, the $MLLR$ test is performed.) Stage one can be performed very quickly and can give excellent preliminary results. Stage two takes longer but yields better power. The results of stage one can be used to estimate the running time of stage two. PREST also identifies Mendelian errors through examination of every father-mother-offspring trio. However, the presence of Mendelian errors is not directly taken into account in the hypothesis tests for detection of pedigree errors. More discussions on Mendelian errors are in Section 4.3.

Currently, the relationships in the set $\mathcal{A}$ for the $MLLR$ test are MZ-twin, parent-offspring, full-sib, half-sib-plus-first-cousin, half-sib, grandparent-grandchild, avuncular, first-cousin, half-avuncular, half-first-cousin and unrelated. (The test statistic $MLLR$ would be the maximum of the log-likelihood over all relationships from this set, excluding the null relationship, minus the null log-likelihood.) PREST uses a default value of 10^5 replicates for the calculation of the empirical p-value of the $MLLR$ test. For the cases of parent-offspring and MZ-twin pairs, PREST uses a default value of $\epsilon = .01$ for the genotyping error rate to calculate the likelihood as described in Section 2.2.3. These default values can be changed by the user.

If the null relationship for a pair, according to the pedigree, is rejected, it is of interest to know what relationship(s) is compatible with the observed genotype data for the pair. To determine this, we have developed a program, ALTERTEST ("ALTERTEST" stands for Alternative Test). For any pair of individuals chosen by the user, ALTERTEST will perform the $EIBD$, $AIBS$, IBS and $MLLR$ tests with user-specified null hypothesis that is different from the one specified by the pedigree. (In contrast, PREST automatically chooses the relationship specified by the pedigree as the null hypothesis for the tests.) Moreover, ALTERTEST allows more than one null hypothesis to be specified for each pair. The purpose of ALTERTEST is to allow the user to convert the tests and construct a confidence set of relationships that are consistent with the observed genotype data. Although ALTERTEST was first constructed as a complement to PREST, it can be used independently of PREST. If one has a null hypothesis in mind for some specific relative pair, one can run ALTERTEST alone. In other cases, it is helpful to run PREST first, so that one can identify questionable pairs, propose some likely relationships for the pairs based on the estimate of $\mathbf{p}$, and then one can apply ALTERTEST with the proposed relationships as the null hypotheses for the pairs. ALTERTEST allows the 11 relationships listed in the previous paragraph as the null hypotheses for all the tests and as the elements in $\mathcal{A}$ for the $MLLR$ test.

Chapter 4

Application and Discussion

4.1 The GAW 11 COGA data

The GAW 11 COGA data are collected for the purpose of mapping genes for susceptibility to alcohol dependence and related phenotypes (Begleiter *et al*. 1999). The data consist of 105 pedigrees, generally 3- or 4-generation, with 1214 individuals in total. The genome-screen includes 296 markers, of which 285 are autosomal markers that are used in our analysis. Among the 1214 individuals, 992 are genotyped but with missing data at some markers. Most of the individuals have > 250 markers typed, but some individuals have as few as 37 markers typed. The average inter-marker recombination fraction is about 0.13. Allele frequencies, estimated with the USER M13 program (Boehnke 1991), are distributed with the data, as are marker order and distances estimated with the CRIMAP program (Lander and Green 1987).

We analyze this data set using methods described in Chapter 3. Among the 5500 typed pairs (no restriction on the types of relationships), we identify and test 5381 pairs that fit into the relationships considered in $\mathcal{A}$. These pairs are divided into 1037 parent-offspring, 1283 full-sib, 79 half-sib, 171 grandparent-grandchild, 942 avuncular, 350 first-cousin, 71 half-avuncular, 40 half-first-cousin and 1407 unrelated pairs. The majority of these pairs have $>$ 200 markers typed in common. The minimum number of shared typed markers is 25. Among the pairs with at least 100 markers typed, there are 78 significant pairs at level 10^{-5}, which corresponds to a level of 0.05 after Bonferroni correction. These 78 pairs occur in 11 pedigrees. Here we discuss only two examples.

The first example illustrates a quite common case of pedigree errors. The pedigree considered is shown in Figure 4.1. This is a family with a sibship of size 3. The starred individuals (the parents) are not genotyped. Among the three sibs, a particular individual 5, is detected to have relationship misfit with each of his two putative full sibs, individuals 3 and 4. The estimates of $\mathbf{p}$ suggested that 5 could be half sibs with 3 and 4. The proposed half-sib relationships, for pair 3 and 4 and pair 4 and 5, are not rejected. The corresponding test results are given in Table 4.1.

The second example is perhaps more interesting, from the point of view of the methods developed here. In this example, the apparently misspecified relationships could not have been detected by methods that check only full-sib or half-sib pairs. Consider the family shown in Figure 4.2. Here we have removed extraneous individuals from the pedigree and changed the sexes of some individuals to provide an extra level of confidentiality for the family. The starred individuals are not typed, while all other individuals have genotype data at $>$ 250 markers. In this pedigree, significant relationship misfit is detected for the reported first-cousin pair, individuals 15 and 16. The empirical p-value of the $MLRT$ test is 0 assessed based on simulation with 10^6 replicates. The estimate of $\mathbf{p} = (p_0, p_1, p_2)$ is $(.281, .555, .164)$, which is between

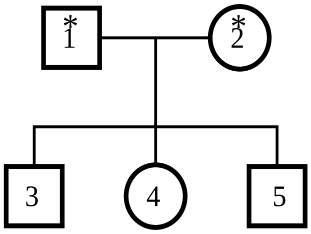

Figure 4.1: Pedigree for the first example of pedigree errors in the COGA data

The starred individuals are untyped, and all others are genotyped at >200 markers.

Pair [a]	N_m [b]	R_o [c]	p-value [d]	$\hat{\mathbf{p}} = (\hat{p_0}, \hat{p_1}, \hat{p_2})$	R_p [e]	p-value [f]
3 4	230	f.sib	.56	(.238 .525 .237)		
3 5	234	f.sib	0	(.542 .449 .009)	h.sib	.42
4 5	237	f.sib	0	(.463 .513 .025)	h.sib	.69

Table 4.1: Testing results for the first example

[a] Individuals are labeled as in Figure 4.1. [b] The number of markers typed in both individuals. [c] The null relationship specified by the pedigree. [d] The empirical p-value (10^6 or 10^7 replicates) of the $MLLR$ test of $H_o : R_o$. [e] The proposed relationship suggested by $\hat{\mathbf{p}}$. [f] The empirical p-value (10^5 replicates) of the $MLLR$ test of $H_o : R_p$.

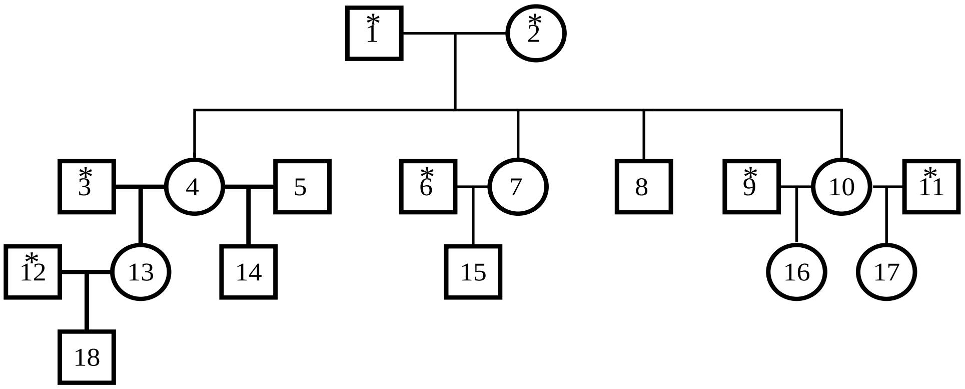

Figure 4.2: Pedigree for the second example of pedigree errors in the COGA data

The starred individuals are untyped, and all others are genotyped at > 250 markers.

half and full sibs. There is no relationship misfit between individuals 7 and 10, who are the mothers of the cousins. Furthermore, there is no relationship misfit detected between 7 and 15, between 10 and 16, between 16 and 17, between 8 and 16, etc. One possible explanation consistent with the data is that individuals 6 and 9, who are fathers of the cousins, are the same person, or, perhaps MZ twins. In that case, individuals 15 and 16 would have the half-sib-plus-first-cousin relationship shown in Figure 2.6, while all other relationships in the pedigree would be preserved. Note that the null IBD probability distribution, (p_0, p_1, p_2), for a half-sib-plus-first-cousin pair is $(.375, .5, .125)$. When then test the half-sib-plus-first-cousin relationship for individuals 15 and 16, the empirical p-value of the $MLRT$ test is 0.266. Since neither 6 nor 9 is typed, we are unable to directly test whether 6 and 9 are the same person (or MZ twins).

4.2 The Hutterite data

The Hutterite data are collected by Dr. Carole Ober and colleagues at the University of Chicago. The data consist of a single pedigree that has 13 generations and 1623 individuals. These individuals are descendants of 64 ancestors who migrated to what is now South Dakota from Europe during the 1870s. They live on communal farms and individuals are typically related through multiple lines of descent. The genome-screen (excluding sex chromosomes) includes 365 markers, with average intermarker recombination fraction of about 0.1. 806 individuals are genotyped, among which most are typed at > 300 markers, but a few individuals are typed at < 5 markers.

This is an extremely complex and highly inbred pedigree. All individuals are related to each other in complicated ways, and no pairs fit into the 11 outbred relationships that we have implemented for the $MLLR$ test. In fact, it is computationally infeasible to calculate the likelihood for pairs in the Hutterite pedigree, therefore to perform the $MLLR$ test. Instead, we propose a graphical method that can be summarized as follows: the first step is to calculate, for each pair of individuals, the probability distribution, $(\Delta_1, ..., \Delta_9)$, of the nine condensed identity states, $(S1, S2, ..., S9)$ illustrated in Figure 2.2. The kinship coefficient Φ of the pair would be $\Delta_1 + (\Delta_3 + \Delta_5 + \Delta_7)/2 + \Delta_8/4$. The second step is to calculate the statistic $EIBD$ as described in Appendix C. Note that states $(S1, S2, ..., S9)$ are assigned to have $(4, 0, 2, 0, 2, 0, 2, 1, 0)$ alleles shared IBD to ensure $E_{R_o}[EIBD] = 4\Phi$. We do not calculate the variance of $EIBD$ because of the computational difficulties due to the complexity of the relationships in the Hutterites. The last step is to plot the observed statistic $EIBD$ for each pair vs. the kinship coefficient Φ for that pair and look for apparent outliers in the graph. One could also calculate statistics $AIBS$ and IBS, and their null means as described in Appendix C. Then, one could plot the observed statistics vs. the null means and look for apparent outliers. However, the computation of the null means of $AIBS$ and IBS is intensive for pairs in the Hutterites. One could plot $AIBS$ or IBS vs. the kinship coefficient to look for outliers. Our analysis in the Hutterites suggests that this approach works well. (The plot based on $AIBS$ is very close to that based on $EIBD$. The plot based on IBS is similar to that based on $EIBD$, but the points on the former are much more dispersing than those on the latter. This is expected, because the IBS statistic has substantially higher null variance than the $EIBD$ and ABS tests.)

To calculate $(\Delta_1, ..., \Delta_9)$ for an inbred pair, Karigl (1981) proposes a recursive algorithm. Although this algorithm works well for moderate-size inbred pedigrees, it is computationally difficult in the complex Hutterite pedigree. Abney, McPeek and Ober (2000) modify the Karigl algorithm using a new computational strategy, so that the calculation of $(\Delta_1, ..., \Delta_9)$ for pairs in the Hutterites becomes feasible. Based on the results of Abney, McPeek and Ober (2000), we obtain 236,597 pairs with > 50 markers typed in common, along with their probability distributions, $(\Delta_1, ..., \Delta_9)$. We then calculate the statistic $EIBD$ and the kinship coefficient Φ for each pair, and plot $EIBD$ vs. Φ. in Figure 4.3.

We find four obvious MZ-twin pairs or duplicated samples (marked with diamonds in Figure 4.3), with all or nearly all the markers identical. We also observe that one individual, denoted as 9 in Figure 4.4, has a number of relationship misfits (marked with x's in Figure 4.3). Figure 4.4 is a partial pedigree showing the position of 9 relative to some intermediate family members in the Hutterites. The starred indi-

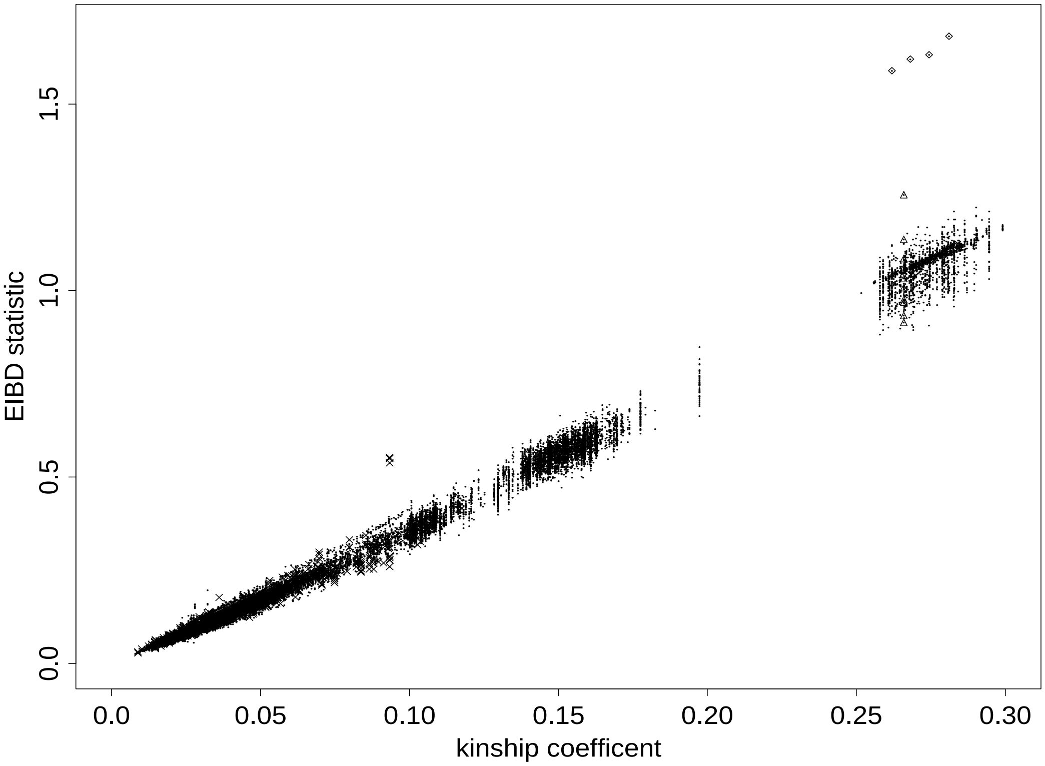

Figure 4.3: *EIBD* vs. kinship coefficient for pairs in the Hutterite data

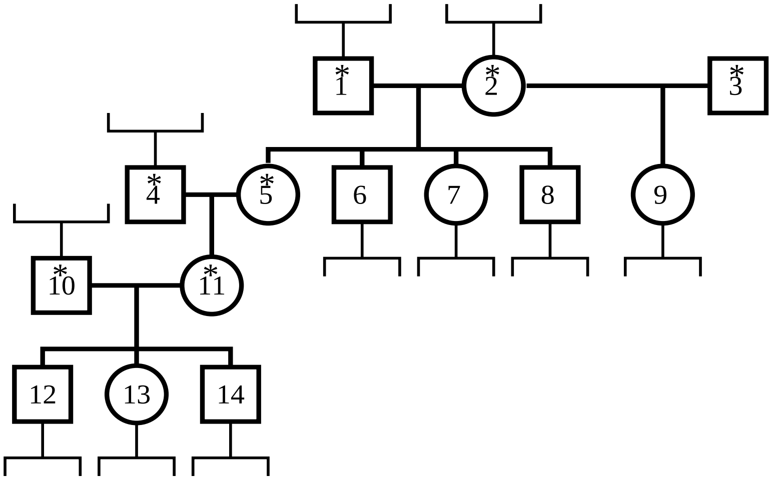

Figure 4.4: A partial pedigree for individual 9 in the Hutterite data

The starred individuals are untyped, and all others are genotyped at > 300 markers.

viduals are not typed, and all other individuals are typed at > 300 markers. Based on the data, 9 shows a large amount of oversharing with individuals 12, 13 and 14, compared to what would be expected based on the pedigree, i.e. kinship coefficient Φ. The estimates of $\mathbf{p} = (p_0, p_1, p_2)$ between 9 and each of the three individuals 12, 13 and 14 are all about $(.008, .992, .000)$. In fact, at almost every marker, 9 shares at least 1 allele IBS with 12, 13 and 14. This could be explained by the possibilities that 9 and 11 are either the same person or MZ twins. There is also one inbred sib pair (marked with a triangle in Figure 4.3) that shows a large amount of oversharing. This pair is from an inbred sibship of size 5, and none of the other 9 pairwise inbred sib pairs (also marked with triangles) show oversharing. The observed oversharing could be due to chance.

4.3 Discussion

Genome-screen data collected for linkage studies can provide considerable power to detect pedigree errors. We have developed a variety of statistical tools for both error detection and relationship estimation. Our methods can be applied to a wide range of pedigree types, from general outbred pedigrees to complex inbred pedigrees. We have implemented all the methods as freely available software, and we have successfully detected a number of misspecified pairs in several data sets collected for linkage studies.

In order to extend the likelihood calculations of Göring and Ott (1997) and Boehnke and Cox (1997) to more general pairwise relationships for which the IBD processes are not Markov, we derive augmented processes that contain the information needed beyond IBD status in order to make the process Markov. Using these augmented Markov processes as the basis of our likelihood calculation, we propose the $MLLR$ test, for which significance is assessed by simulation. Extensions of the likelihood calculation that take into account interference, inbreeding and genotyping errors are also discussed. We then extend the IBS test of Ehm and Wagner (1998) to more general relative pairs. The implementation of the IBS test is much simpler computationally than that of the $MLLR$ test, but the IBS test loses power by not taking into account chance sharing. As a compromise between the two, we propose two new tests based on $EIBD$ and $AIBS$. The power of the $EIBD$ and $AIBS$ tests is not too much lower than that of the $MLLR$ test, while they maintain the desirable features of the IBS test.

Among the four tests considered, the $MLLR$ test has the highest power to detect misspecified relationships. However, it requires one to specify a set of alternative relationships and augment separate Markov process for each as well as for the null relationship, while the $EIBD$, $AIBS$ and IBS tests do not need specification of any alternative. Furthermore, assessment of significance for $MLLR$ requires computationally intensive simulation, while the normal approximation can be used for $EIBD$, $AIBS$ and IBS. For data in which there is a large number of pairs to be tested, e.g. the GAW 11 COGA data, we propose two screening procedures, as described in Appendix, of which one has been implemented in our software.

One way to make the application of $MLLR$ feasible for a wider class of relationships is to use the Markov approximation to the likelihood proposed in Section 2.2.5. In that case, rather than construct an augmented Markov process for each null and alternative relationship considered, one need only calculate the one-step conditional IBD probabilities $\Pr(D_{m2} = j | D_{m1} = i)$, where D_m is the number of alleles shared IBD by a given pair at locus m. Our results indicated that at least for the avuncular, first-cousin, half-avuncular, half-first-cousin and half-sib-plus-first-cousin relationships, this approximation is adequate for relationship testing.

Although the current implementation of $MLLR$ does not include likelihood calculation for inbred relationships, it is still possible to detect instances of relative marriage by use of our methods. If putatively unrelated parents are in fact related, and if they are both typed, our methods have some power to detect their relationship, with greater power for closer relationships (Sun, Abney and McPeek, in press). Note that in this case, offspring genotypes provide no additional information on the relatedness of the parents, conditional on the parental genotypes and the correct paternity/maternity classification for the offspring. However, if both parents are untyped, there can be low power to detect slight inbreeding in a sib pair. Power could potentially be increased by the development of specially designed methods to detect inbreeding in a sibship with one or both parents untyped.

For the test statistics $EIBD$, $AIBS$ and IBS, one could consider trying to increase power by weighting markers differently depending on their location in the genome, giving isolated markers more weight than those in densely mapped regions, because correlation between markers is a decreasing function of distance. Our preliminary work on optimal weights shows that, at least in the complete data case, the increase in power tends to be small (results not shown).

Currently, our methods do not use data from the sex chromosomes. Epstein *et al.* (2000) show that X-linked data can provide some information to discriminate among certain sex-specific types of second-degree relationships.

For a single hypothesis test, it is conventional to use a significance level of 0.05 or 0.01. For data such as the COGA data, there are a large number of pairs to be tested. This creates a problem of multiple comparisons, e.g. even if the null hypothesis is true in all cases, some fraction of the pairs would be expected to have p-values below 0.05. To control the false positive rate to be, say 0.05, one could apply the Bonferroni correction and use $0.05/N$ as the threshold, where N is the number of hypothesis tests performed. In our study, pairs in a pedigree are not independent. Thus, this correction is rather conservative. In the case where Mendelian errors have been cleaned, one may be able to reduce the size of N by not counting the parent-offspring pairs as discussed in Sun, Abney and McPeek (in press).

When a particular pair is observed to have significant deviation from its null relationship, the pattern among other relatives can often confirm and point to a likely explanation for the finding. For instance, consider a family with a sibship of size 3 (ID 1, ID 2 and ID 3) with parents untyped. If ID 1 are half-sib to the other two, then one would expect to detect an error in two different pairs (ID 1, ID 2) and (ID 1, ID 3), and not expect to see a relationship misfit in pair (ID 2, ID 3). Example one from the COGA data discussed in Section 4.1 fits exactly this case. Thus, when an apparent error is detected, by considering the pattern of results among multiple pairs from the same pedigree, one many be able to distinguish a true relationship error from chance rejection of the null. Ideally, the methods presented in this paper should be extended to simultaneous inference on a number of relatives, e.g., an entire sibship could be considered simultaneously in a single likelihood analysis, rather than separate consideration of pairwise relationships. However, as discussed before, to quickly identify the problematic individuals in a pedigree and effectively propose possible alternatives, a pairwise approach is useful. In some cases such as the Hutterites, joint analysis may not be computationally feasible.

Errors that are incompatible with Mendelian inheritance are called Mendelian errors, e.g. a father has genotype ($a1$ $a2$) and a mother has genotype ($a2$ $a2$), while one of their offspring has genotype ($a2$ $a3$). The presence of Mendelian errors can be the result of genotyping errors, but it may also be an indication of pedigree errors, especially if the level of Mendelian errors is high. Mendelian errors can be detected by existing software such as PedCheck (O'Connell and Weeks 1998). PREST is developed specifically for the purpose of detecting those misspecified relationships that are not already detected through routine checks for Mendelian errors. Thus, the program does not do an exhaustive search for Mendelian errors. It does, however, check for Mendelian errors through examination of every mother-father-child trio.

Currently, Mendelian errors in a pedigree are not directly taken into account in the hypothesis tests for pairs from that pedigree. However, their presence or absence and overall level can provide important clues to pedigree errors and to the understanding of their likely causes. Thus, to detect likely pedigree errors, it is useful to combine the information on Mendelian errors with the hypothesis test results.

Part II

Identification of Polymorphisms That Explain a Linkage Result

Chapter 5

Introduction

5.1 Genetic mapping studies

To identify genetic variation affecting susceptibility to complex disease, there are generally sequential stages involved, from coarse, genome-wide linkage analysis, to fine mapping, and then to positional cloning. Linkage analysis typically looks for genetic markers, among a large number of markers typed throughout the genome, at which there is a significant deviation of the IBD allele-sharing by affected relatives from what expected under the null hypothesis. (For the definitions of linkage and IBD, and for further genetics background, see Sections 1.3 and 1.4.) The initial detection of linkage takes the form of a hypothesis test, in which the null hypothesis is that there is no linkage between the marker site to be tested and a susceptibility locus for the trait. Although this approach is successful in performing a genome-wide search, i.e. knowing approximately the location of the gene, it rarely provides map resolution finer than 1 centiMorgan (cM) which corresponds to roughly 1 million DNA base pairs (bp).

Fine mapping utilizing linkage disequilibrium (LD) is often used to further narrow down the region. LD refers to the lack of independence of alleles at loci on a haplotype randomly sampled from a population. Imagine the disease originated, say 50 generations ago, from a single mutation at a locus in a halplotype with a unique set of marker alleles. These marker alleles are likely to be co-inherited with the disease allele by the affected individuals. Because recombination may occur among these loci during each meiosis, as the haplotype is transmitted into the following generations, the chance that any given characteristic allele remains in the same haplotype decreases with increasing genetic distance from the disease allele and with increasing number of generations. When one collects data after 50 generations, alleles at many of these marker loci may be independent of alleles at the disease locus, at the population level, and only a very small region around the disease locus is likely to retain the characteristic haplotype and remain in LD with the disease locus. Therefore, for loci that are linked, unless the chance of recombination among them is extremely small, they will not necessarily be in LD. This distinction between linkage and LD is important in developing mapping strategies. Because linkage tends to operate over greater chromosomal distances than LD, a sequential mapping strategy in which initial linkage analysis is followed by LD mapping, is often used. The map resolution achieved by LD mapping is generally much higher than that obtained through linkage analysis. Linkage disequilibrium between a pair of loci, 1 and 2, can be summarized by $D' = D/D_{max}$, if $D > 0$, or $= D/D_{min}$, if $D < 0$, where D is the disequilibrium parameter defined as $D = p_{ab} - p_a p_b$. ($D = 0$ implies linkage equilibrium.) $D_{max} = min(p_a(1-p_b),(1-p_a)p_b)$, $D_{min} = max(-p_a p_a, -(1-p_a)(1-p_b))$, p_{ab} is the probability of allele a at locus 1 and allele b at locus 2 occurring on the same chromosome strand, i.e. probability of haplotype ab, and p_a and p_b are the frequencies of alleles a and b.

After a susceptibility locus has been localized to a rather small region by linkage analysis and fine mapping, one may be able to identify one or several genes within that region. Even if only one gene lies in the region, a large number of DNA polymorphic sites may still exist. The goal of positional cloning is to identify the causal sites among all the polymorphic sites in a gene. Ultimately, only biological study can verify that certain genetic variation has the consequence of increasing susceptibility to disease. However, statistical analysis of the available data can provide guidance on which variants merit the next level of biological study. Many statistical methods have been developed for the first two stages of the process, i.e. linkage analysis and fine mapping. We focus on the third stage, and we describe here a new statistical approach to guide positional cloning.

5.2 Positional cloning

Suppose that many polymorphic sites have been identified and genotyped in a region showing strong linkage with a disease or trait. We assume that these sites are all tightly linked and that they may be in linkage disequilibrium with each other and with the susceptibility locus. We would ideally like to determine which site or combination of sites in the region influences susceptibility to the trait. To accomplish this, we need to distinguish the actual causal site from other sites that are merely tightly linked or in linkage disequilibrium with the causal site.

Previous work on statistical methods for positional cloning of quantitative traits include Fulker *et al.* (1999), Cardon and Abecasis (2000), Soria *et al.* (2000), Siegmund *et al.* (in press) and Blangero *et al.* (in press). The approach of Fulker *et al.* (1999) is developed in the context of a variance components approach to combined linkage and association analysis of quantitative traits in sib pairs. Fulker *et al.* (1999) point out that testing linkage while simultaneously modeling association would provide a test of whether the putative quantitative trait locus (QTL) is a candidate or is merely in disequilibrium with a trait locus. This idea is further developed by Cardon and Abecasis (2000), who also consider the implications for the possible range of allele frequencies for the candidate locus. In a similar context of quantitative trait analysis, Soria *et al.* (2000) note that if there is only one causal variant in a region, then linkage analysis that is performed conditional on the measured genotypes should yield no evidence for linkage. They use this idea to argue that the prothrombin G20210A mutation affects the function of the prothrombin gene. A similar approach is used in simulation studies by Siegmund *et al.* (in press). Blangero *et al.* (2000) propose a Bayesian model selection/averaging method for positional cloning of quantitative traits. They extend the classical variance component model and utilize Bayesian methods to estimate the posterior probability that each polymorphic site is the variant that is responsible for the variation present in the phenotype. They consider only additive effects at the hypothesized causal site because of the computational difficulties. They apply their quantitative nucleotide analysis to the GAW12 simulated data, in which a single SNP in a gene influences the quantitative trait of interest. The true SNP is correctly inferred as it has the highest posterior probability among all the 18 SNPs considered.

For qualitative traits, a statistical method for positional cloning is proposed by

Horikawa *et al.* (2000). They suggest a modified association study in which they examine not only the differences in allele frequencies between controls and cases, but also how the evidence for linkage is partitioned in pairs defined by the genotype at the SNP to be tested. They observe that, under the null hypothesis of no association between a particular SNP and the trait, if affected sib pairs are classified according to the genotype at the SNP, the observed lod score should be divided into each group proportional to what is expected for each genotype category under the null hypothesis. They perform simulation to assess the p-value of the observed lod score in a group in which both sibs have the at-risk genotype(s) at the SNP to be tested, and they identify a SNP (UCSNP-43) that shows significant association with the evidence for linkage with type 2 diabetes.

Some methods originally developed for other purposes are conceptually similar to those described above. For example, Greenberg (1993) suggest a partitioned association-linkage test, further developed by Hodge (1993), in which affected sib pairs are partitioned on the basis of the presence or absence of an associated allele in the index case, and the IBD sharing is assessed separately in the affected sib pairs where the index case does and does not have the associated allele. The test is originally proposed as a way of distinguishing loci necessary for the development of a disease from those that merely increase susceptibility, but is similar to the approaches described in Horikawa *et al.* (2000) for identifying variants showing association with the evidence for linkage.

Our method is designed for the problem of positional cloning studies for qualitative traits. Our approach to this problem is to identify the polymorphisms whose genotypes could fully explain, in the statistical sense, the observed linkage to the region. We frame the question as a hypothesis test. We focus on the case in which we assume that there is only one causal polymorphic site in the region segregating in the study population, with allelic heterogeneity allowed. (We also discuss an extension to multiple tightly-linked polymorphic sites influencing the trait.) Under the single-site assumption, for a given polymorphic site in the region, the null hypothesis is that the site considered is the sole cause of linkage to the region. We observe that, under this null hypothesis, the conditional distribution of IBD sharing among the affected relatives, in the region, given their genotypes at the putative causal locus, does not depend on the genetic model for the trait. A departure from the null hypothesis implies that the hypothesized site is not the sole cause of linkage to the region. Such a hypothesis test can be performed on each of the polymorphic sites typed in the region of interest. A confidence set for the true casual site can be constructed by inverting the hypothesis test, that is, by including in the confidence set all the sites that are not rejected by the hypothesis test (including those not tested). The results of this approach provide information that is different from that provided by tests of linkage or association.

To implement our approach, we focus on the sib-pair study design with single nucleotide polymorphisms typed in the region of interest, and we consider test statistics that are variations on the usual allele-sharing methods used for linkage studies. Our approach does not require specification of mode of inheritance at the putative causal polymorphism. Moreover, our method allows an arbitrary amount of epistasis with other unlinked contributory loci, as well as correlated environmental effects within families, and gene-environment interaction. We extend our method to larger sibships, and we apply it to a data set developed in the context of a positional cloning study

(Horikawa *et al* 2000). Through both simulation studies and data analysis, we find that we have power to reject sites that do not, on their own, explain the evidence for linkage, even when these sites are both tightly linked and strongly associated with a susceptibility locus.

Chapter 6

A Novel Approach for Positional Cloning

6.1 Conditional distribution of IBD sharing

We first consider the case of sib pairs sampled at random from a population, without regard to their phenotypes. (For simplicity, we use sib pairs to denote full-sib pairs.) For this case, we derive the distribution of IBD sharing by a sib pair at a particular SNP, conditional on the sibs' genotypes at that SNP. We then consider the case in which affected sib pairs are sampled. We show that, under the null hypothesis that a particular SNP is the sole cause of linkage to the region, the distribution of IBD sharing by an affected sib pair, conditional on the sibs' genotypes at that SNP, is the same as in the case of random sib pairs. We argue that this is true regardless of the mode of inheritance and even in the presence of epistasis with unlinked loci, correlated environmental effects within families, and gene-environment interaction. This result allows us to test for deviation from the null conditional distribution of IBD sharing by affected sib pairs and to construct a confidence set of polymorphisms that could explain the observed linkage to the region.

6.1.1 For random sib pairs

Consider the case in which sib pairs are drawn at random from a population, regardless of their phenotypes. We derive the conditional distribution of IBD sharing by such a sib pair at a particular SNP, given the sibs' genotypes at that SNP. Denote by 1 and 2 the two alleles of the SNP, and let $g_1 = (1\ 1\ 1\ 1)$, $g_2 = (1\ 1\ 1\ 2)$, $g_3 = (1\ 1\ 2\ 2)$, $g_4 = (1\ 2\ 1\ 2)$, $g_5 = (1\ 2\ 2\ 2)$ and $g_6 = (2\ 2\ 2\ 2)$ be the 6 possible genotype configurations for the sib pair at the SNP, where the first two integers represent the genotype of one sib, the last two integers represent the genotype of the other sib, and where we consider two sib-pair genotype configurations to be equivalent if they are the same up to interchange of the two sibs and/or interchange of the two alleles of either sib. To complete the notation, let f be the frequency of allele 1, let G be the random variable representing the sibs' genotype configuration at the SNP and taking values in $\{g_1, g_2, ..., g_6\}$, and let D be the number of alleles shared IBD by the pair at the SNP locus. Table 1 gives the conditional distribution of $P(D|G)$. The following equation illustrates the calculation for the case when $G = g_1 = (1\ 1\ 1\ 1)$, and $D = 1$.

$$\begin{aligned} P(D=1 \mid G=(1\ 1\ 1\ 1)) &= \frac{P(D=1, G=(1\ 1\ 1\ 1))}{P(G=(1\ 1\ 1\ 1))} \\ &= \frac{P(G=(1\ 1\ 1\ 1)|D=1)P(D=1)}{\sum_{j=0,1,2} P(G=(1\ 1\ 1\ 1)|D=j)P(D=j)} \\ &= \frac{f^3\frac{1}{2}}{f^4\frac{1}{4}+f^3\frac{1}{2}+f^2\frac{1}{4}} = \frac{2f}{(1+f)^2}. \end{aligned}$$

Note that $P(D)$ depends on the relationship of the two individuals and $P(G|D)$ is calculated under the assumption of Hardy-Weinberg equilibrium (HWE). HWE means the frequency of the genotype at a locus for an individual, $(a_i\ \ a_j)$, depends only on the frequencies of the alleles, i.e. $P(a_i \;\; a_j) = p_{a_i}^2$ if $i = j$, and $P(a_i \;\; a_j) = 2p_{a_i}p_{a_j}$ if $i \neq j$, where p_{a_i} and p_{a_j} are frequencies of a_i and a_j alleles. The computation of $P(G|D)$ for a pair of outbred individuals appears in Thompson (1975).

G	D		
	0	1	2
(1 1 1 1)	$\frac{f^2}{(1+f)^2}$	$\frac{2f}{(1+f)^2}$	$\frac{1}{(1+f)^2}$
(1 1 1 2)	$\frac{f}{1+f}$	$\frac{1}{1+f}$	0
(1 1 2 2)	1	0	0
(1 2 1 2)	$\frac{f(1-f)}{1+f(1-f)}$	$\frac{1}{2(1+f(1-f))}$	$\frac{1}{2(1+f(1-f))}$
(1 2 2 2)	$\frac{1-f}{1+(1-f)}$	$\frac{1}{1+(1-f)}$	0
(2 2 2 2)	$\frac{(1-f)^2}{(1+(1-f))^2}$	$\frac{2(1-f)}{(1+(1-f))^2}$	$\frac{1}{(1+(1-f))^2}$

Table 6.1: $P(D|G)$ for a sib pair with SNP data

The conditional distribution of $P(D|G)$, where D is the number of alleles shared IBD by a sib pair at a particular SNP, G is the sibs' genotype configuration at that SNP, and f is the frequency of allele 1 in the population.

6.1.2 For affected sib pairs

We now consider the case in which affected sib pairs are drawn at random from a population. We show that, under the null hypothesis H_o that a particular SNP is the sole cause of linkage to the region, the conditional distribution of IBD sharing by an affected sib pair, given the sibs' genotype configuration at that SNP, is the same as in the previous case, i.e. Table 6.1. To show this, we first argue that the following equation holds regardless of the mode of inheritance:

$$P_{H_O}(\text{both affected}|D, G) = P_{H_O}(\text{both affected}|G).$$

That is, given the genotype data at the sole causal site in the region for an affected sib pair, the event that both sibs are affected by the trait is a Bernoulli trial with probability depending only on the observed genotypes, independent of the sharing at that location, as long as the other causal loci are not linked to the region. The above equation implies the following equation which states that the conditional distribution of IBD sharing by randomly-sampled affected sib pairs is the same as that for randomly-sampled sib pairs regardless of their phenotypes.

$$\begin{aligned} P_{H_O}(D|G, \text{both affected}) &= \frac{P_{H_O}(\text{both affected}|D, G)P_{H_O}(D, G)}{P_{H_O}(\text{both affected}|G)P_{H_O}(G)} \\ &= \frac{P_{H_O}(D, G)}{P_{H_O}(G)} = P_{H_O}(D|G) = P(D|G), \end{aligned}$$

where $P_{H_O}(D|G) = P(D|G)$ because neither expression contains phenotype information of the sibs.

6.2 Hypothesis testing and confidence set construction

In the previous subsection, we have shown that, under the null hypothesis that a particular SNP is the sole cause of linkage to the region, the conditional distribution of IBD sharing by an affected sib pair, in the region, given the sibs' genotype configuration at that SNP, can be derived without specification of the mode of inheritance and is given by Table 6.1. For any SNP typed in the region, to test the null hypothesis

$$H_o : \text{the SNP is the sole causal site in the region,}$$

we could construct a test that is a variation on whatever method was used to detect linkage initially. (However, note that our test is not a test for linkage. In fact, we expect that all the polymorphisms in the region will be tightly linked to the susceptibility locus.) For instance, suppose linkage was initially detected using an allele-sharing method with a given sharing statistic S which measures the IBD sharing D among a set of affected relatives, e.g. $S = S_{pairs}$ which is the sum of the number of alleles shared IBD over all possible pairs of the affected relatives. The null distribution of $P(D|G)$ derived in the previous subsection allows us to calculate the null conditional mean and null conditional standard deviation of S, $\mu_G = E_{H_o}[S|G]$ and $\sigma_G = \sqrt{Var_{H_o}(S|G)}$, where G is the sibs' genotype configuration at the SNP, and H_o is the null hypothesis that the SNP is the sole causal site in the region. For an affected sib pair, Table 6.2 gives μ_G and σ_G, when S_{pairs} is used, for each of the 6 genotype configurations. To test our null hypothesis H_o, we could use a variation on the NPL score statistic of Kruglyak *et al.* (1996), the linear likelihood of Whittemore (1996) and Kong and Cox (1997), or the exponential likelihood of Kong and Cox (1997). Consider the usual tests for detection of linkage by these methods, and let H_o' be the null hypothesis of no linkage, let μ and σ be the unconditional mean and standard deviation of S under H_o', $\mu = E_{H_o'}[S]$ and $\sigma = \sqrt{Var_{H_o'}(S)}$, and let $Z' = (S - \mu)/\sigma$ be the standardized version of S for a particular family, for the usual test of linkage. To modify any of these linkage methods to test our null hypothesis H_o, we replace $Z' = (S - \mu)/\sigma$ by $Z^G = (S - \mu_G)/\sigma_G$ for each family. (If $\sigma_G = 0$,

G	μ_G	σ_G
(1 1 1 1)	$\frac{2}{1+f}$	$\frac{\sqrt{2f}}{1+f}$
(1 1 1 2)	$\frac{1}{1+f}$	$\frac{\sqrt{f}}{1+f}$
(1 1 2 2)	0	0
(1 2 1 2)	$\frac{3}{2(1+f(1-f))}$	$\frac{\sqrt{1+10f(1-f)}}{2(1+f(1-f))}$
(1 2 2 2)	$\frac{1}{1+(1-f)}$	$\frac{\sqrt{(1-f)}}{1+(1-f)}$
(2 2 2 2)	$\frac{2}{1+(1-f)}$	$\frac{\sqrt{2(1-f)}}{1+(1-f)}$

Table 6.2: μ_G and σ_G for an affected sib pair with SNP data

The null conditional mean, $\mu_G = E_{H_o}[S_{pairs}|G]$, and the null conditional standard deviation, $\sigma_G = \sqrt{Var_{H_o}(S_{pairs}|G)}$, of the sharing statistic S_{pairs} for an affected sib pair, given the sibs' genotype configuration G at a particular SNP, under the null hypothesis H_o that the SNP is the sole causal site in the region. f is the frequency of allele 1 in the population.

then $Z^G = 0$.) Note that while μ and σ depend only on the relationship among the affecteds, μ_G and σ_G also depend on G, the genotype configuration for the affecteds at the SNP.

Given n affected sib pairs, let D_i be the number of alleles shared IBD by the ith sib pair, let S_i be the sharing statistic for the i_{th} pair, let G_i be the observed genotype configuration for the ith pair at the SNP to be tested, and let

$$Z_i^G = (S_i - \mu_{G_i})/\sigma_{G_i} = (S_i - E_{H_o}[S|G_i])/\sqrt{Var_{H_o}(S|G_i)}$$

be our new, conditional, standardized version of S_i. We could then consider the test statistic T_1 (which is in a form analogous to that of the NPL score statistic for testing linkage) for our null hypothesis:

$$T_1 = \frac{\sum_{i=1}^{n} w_i Z_i^G}{\sqrt{\sum_{i=1}^{n} w_i^2}}, \tag{6.1}$$

where w_i is the weighting factor for the i_{th} family. We could also consider the test statistic T_2 (which is in a form analogous to that of the exponential log–likelihood-ratio for testing linkage):

$$T_2 = \text{sign}(\hat{\delta})\sqrt{2[l(\hat{\delta}) - l(0)]}, \tag{6.2}$$

where $l(\delta) - l(0) = \log[\Pi_i c_i(\delta)\exp(\delta w_i Z_i^G)]$, $c_i(\delta) = [\sum_z P_{H_o}(Z_i^G = z|G_i)\exp(\delta w_i z)]^{-1}$ is the renormalization constant, $P_{H_o}(Z_i^G = z|G_i) = P_{H_o}(S_i = z\sigma_{G_i} + \mu_{G_i}|G_i)$, which can be calculated from the information in Tables 6.1 and 6.2 for the case of S_{pairs} in an affected sib pair, and $\hat{\delta}$ maximizes $l(\delta)$ (i.e. it maximizes $l(\delta) - l(0)$). With complete IBD data, the tests based on the statistics in equations (6.1) and (6.2) are equivalent (assuming that exact p-values are used), with the version in equation

(6.1) being easier to calculate. However, with incomplete IBD information or when a small number of large families are sampled and the normal approximation is used to assess significance, the test based on equation (6.2) is preferred (Kong and Cox 1997). The case of incomplete IBD information is discussed in more detail in Section 6.4. Another possible variation would be to use test statistic T_3, which is analogous to the linear log–likelihood-ratio and is of the same form as equation (6.2), but here $l(\delta) - l(0) = \log[\Pi_i(1 + \delta w_i Z_i^G)]$. The test based on this statistic is not equivalent to either of the previous two tests, except asymptotically. To assess significance, one could use simulations to obtain the empirical distribution of the test statistic T_1, T_2 or T_3, conditional on $(G_1, G_2, \ldots, G_n)$. To do this, we simulate D_i conditional on G_i for each i, using the distribution of $P(D|G)$ given in Table 6.1. Alternatively, one could apply a normal approximation to the conditional distribution of T_1, T_2 or T_3. In principle, the test could be two-sided. However, we note that the SNP is assumed to be in a region showing strong linkage with a trait. Therefore, $\hat{\delta} > 0$ is expected if the SNP is not the sole cause of linkage to the region, whereas $\hat{\delta} < 0$ may indicate misspecification of the allele frequency f or violation of the Hardy-Weinberg assumption, which is useful information but is not the alternative of interest. Thus, even if we were to use a two-sided test, we would want to distinguish between these two cases. To construct a confidence set for the true causal site, we perform the corresponding hypothesis test on each of the SNPs typed in the region. A $(1 - \alpha)$ confidence set then includes all the SNPs that are not significant at level α.

Just as for tests of linkage, there are many different possible choices of weighting factor w_i for the ith family when our standardized sharing statistic Z^G is combined across families. The optimal weight for a particular family depends on the amount of information contained in the observed genotype data at the SNP. For instance, the weight for an affected sib pair with genotype configuration $g_3 = (1\ 1\ 2\ 2)$ should be zero, since there is no variation in the IBD sharing given this genotype configuration. In other words, a pair with genotype configuration (1 1 2 2) does not provide any information under our method. For pairs with the other five genotype configurations, one could choose equal weights, $w = 1$, or choose weights that depend on the null conditional variances, such as $w = \sqrt{\sigma_G}$ or $w = \sigma_G$.

We point out that our test is neither a test of linkage nor a test of linkage disequilibrium. A SNP may be tightly linked or in significant linkage disequilibrium with the causal polymorphism, yet still not be able to fully explain the linkage signal observed in the region. In the *NIDDM1* data set of Horikawa *et al.* (2000), SNPs 22, 23, 25, 26, 29 and 38 all show significant linkage and linkage disequilibrium (Horikawa *et al.* 2000), but each is rejected as being the sole cause of linkage to the region (see Section 7.2) Our simulations (see Section 7.1) also show that there are cases in which a false putative causal SNP is both completely linked ($\theta = 0$) and in complete linkage disequilibrium ($|D'| = 1$) with the true causal SNP, and yet our test still has some power to reject the null hypothesis. Of course, if two SNPs are in perfect linkage disequilibrium (i.e. $|D'| = 1$ with the coupled alleles having identical allele frequencies), then they are indistinguishable based on the data and no statistical method can separate them.

6.3 Extension of the method to larger sibships

For the case of more than 2 affected relatives, G would be the genotype configuration among the affecteds in the family, and D would be their IBD configuration (Thompson 1974). For instance, for an affected sib trio with SNP data, there are 10 possible genotype configurations (up to interchange of the three sibs and/or interchange of the two alleles of any sib) and 4 IBD configurations. To calculate the conditional distribution of $P(D|G)$ for an affected sib trio, one needs the conditional distribution of $P(G|D)$ and the marginal distribution of $P(D)$. The conditional distribution of $P(G|D)$ are given in Table 6.3. The marginal distribution of $P(D)$ is $P(D = (12\ 12\ 34)) = 3/16$, $P(D = (12\ 13\ 24)) = 6/16$, $P(D = (12\ 12\ 23)) = 6/16$, and $P(D = (12\ 12\ 12)) = 1/16$. Table 6.4 gives the conditional distribution of $P(D|G)$, and Table 6.5 gives the null conditional mean and null conditional standard deviation of S, when S_{pairs} is used, for each of the 10 genotype configurations. We have derived similar results for sibships with 4-6 affected sibs (results not shown). We have implemented our method for affected sibships of sizes 2-6 and have applied it to the *NIDDM1* data set of Horikawa *et al.* (2000) (see Section 7.2).

6.4 Extension to incomplete IBD data

The extension of our tests to the case of incomplete IBD information is similar to that for the usual allele-sharing tests of linkage. For the usual test of linkage, when the NPL score statistic or the linear likelihood is used with incomplete IBD data, S is replaced by $E_{H_o'}[S|G^{\text{full}}]$, the null expected value of S conditional on

G	IBD configuration D			
	1 2 1 2 3 4	1 2 1 3 2 4	1 2 1 2 2 3	1 2 1 2 1 2
(1 1 1 1 1 1)	f^4	f^4	f^3	f^2
(1 1 1 1 1 2)	$2f^3(1-f)$	$2f^3(1-f)$	$f^2(1-f)$	0
(1 1 1 1 2 2)	$f^2(1-f)^2$	0	0	0
(1 1 1 2 1 2)	$2f^3(1-f)$	$2f^3(1-f)+f^2(1-f)^2$	$f^2(1-f)$	0
(1 1 1 2 2 2)	0	$2f^2(1-f)^2$	0	0
(1 1 2 2 2 2)	$f^2(1-f)^2$	0	0	0
(1 2 1 2 1 2)	$4f^2(1-f)^2$	$2f^2(1-f)^2$	$f(1-f)$	$2f(1-f)$
(1 2 1 2 2 2)	$2f(1-f)^3$	$f^2(1-f)^2+2f(1-f)^3$	$f(1-f)^2$	0
(1 2 2 2 2 2)	$2f(1-f)^3$	$2f(1-f)^3$	$f(1-f)^2$	0
(2 2 2 2 2 2)	$(1-f)^4$	$(1-f)^4$	$(1-f)^3$	$(1-f)^2$

Table 6.3: $P(G|D)$ for a sib trio with SNP data

The conditional distribution, $P(G|D)$, of genotype configuration G at a particular SNP for a sib trio, conditional on the trio's IBD configuration D at that SNP, where f is the frequency of allele 1 in the population.

G	IBD configuration D			
	1 2 1 2 3 4	1 2 1 3 2 4	1 2 1 2 2 3	1 2 1 2 1 2
(1 1 1 1 1 1)	$\frac{3f^2}{(1+3f)^2}$	$\frac{6f^2}{(1+3f)^2}$	$\frac{6f}{(1+3f)^2}$	$\frac{1}{(1+3f)^2}$
(1 1 1 1 1 2)	$\frac{f}{1+3f}$	$\frac{2f}{1+3f}$	$\frac{1}{1+3f}$	0
(1 1 1 1 2 2)	1	0	0	0
(1 1 1 2 1 2)	$\frac{f}{2(1+f)}$	$\frac{1}{2}$	$\frac{1}{2(1+f)}$	0
(1 1 1 2 2 2)	0	1	0	0
(1 1 2 2 2 2)	1	0	0	0
(1 2 1 2 1 2)	$\frac{3}{2}\frac{f(1-f)}{1+3f(1-f)}$	$\frac{3}{2}\frac{f(1-f)}{1+3f(1-f)}$	$\frac{3}{4}\frac{1}{1+3f(1-f)}$	$\frac{1}{4}\frac{1}{1+3f(1-f)}$
(1 2 1 2 2 2)	$\frac{1-f}{2(2-f)}$	$\frac{1}{2}$	$\frac{1}{2(2-f)}$	0
(1 2 2 2 2 2)	$\frac{1-f}{4-3f}$	$\frac{2(1-f)}{4-3f}$	$\frac{1}{4-3f}$	0
(2 2 2 2 2 2)	$\frac{3(1-f)^2}{(4-3f)^2}$	$\frac{6(1-f)^2}{(4-3f)^2}$	$\frac{6(1-f)}{(4-3f)^2}$	$\frac{1}{(4-3f)^2}$

Table 6.4: $P(D|G)$ for an affected sib trio with SNP data

The conditional distribution, $P(D|G)$, of the IBD configuration D for a sib trio at a particular SNP, conditional on the trio's genotype configuration G at that SNP, where f is the frequency of allele 1 in the population.

G	μ_G	σ_G
(1 1 1 1 1 1)	$\frac{6(1+f)}{1+3f}$	$\frac{2\sqrt{6f}}{1+3f}$
(1 1 1 1 1 2)	$\frac{2(2+3f)}{1+3f}$	$\frac{2\sqrt{3f}}{1+3f}$
(1 1 1 1 2 2)	2	0
(1 1 1 2 1 2)	$\frac{3+2f}{1+f}$	$\frac{\sqrt{1+2f}}{1+f}$
(1 1 1 2 2 2)	2	0
(1 1 2 2 2 2)	2	0
(1 2 1 2 1 2)	$\frac{3}{2}\frac{3+4f(1-f)}{1+3f(1-f)}$	$\frac{1}{2}\frac{\sqrt{3+84f(1-f)}}{1+3f(1-f)}$
(1 2 1 2 2 2)	$\frac{5-2f}{2-f}$	$\frac{\sqrt{3-2f}}{2-f}$
(1 2 2 2 2 2)	$\frac{2(5-3f)}{4-3f}$	$\frac{2\sqrt{3(1-f)}}{4-3f}$
(2 2 2 2 2 2)	$\frac{6(2-f)}{4-3f}$	$\frac{2\sqrt{6(1-f)}}{4-3f}$

Table 6.5: μ_G and σ_G for an affected sib trio with SNP data

The null conditional mean, $\mu_G = E_{H_O}[S_{pairs}|G]$, and the null conditional standard deviation, $\sigma_G = \sqrt{Var_{H_O}(S_{pairs}|G)}$, of the sharing statistic S_{pairs} for an affected sib trio, given the trio's genotype configuration G at a particular SNP, under the null hypothesis H_o that the SNP is the sole causal site in the region, where f is the frequency of allele 1 in the population.

G^{full}, the genotype data for all members of the family at all loci at which they are typed. For the usual test of linkage, when the exponential likelihood is used with incomplete IBD data, $\exp(\delta w_i Z_i')$ is replaced by $E_{H_o'}(\exp(\delta w_i Z_i')|G^{\text{full}})$. In the case of the NPL statistic, the above incomplete-data formulation is conservative when the normal approximation is applied, because the variance used to normalize the statistic is too large (Kruglyak *et al.* 1996; Kong and Cox 1997). However, for the linear and exponential likelihoods, these incomplete-data formulations provide an exact likelihood calculation (Kong and Cox 1997). For our test, when G is observed but the IBD information D at this locus is incomplete, the analogous result is that when using T_1 or T_3, S is replaced by $E_{H_o}[S|G^{\text{full}}] = E_{H_o'}[S|G^{\text{full}}]$, where the equality holds because there is no phenotype information on either side of the equation. The analogous result for T_2 is that $\exp(\delta w_i Z_i^G)$ is replaced by $E_{H_o}(\exp(\delta w_i Z_i^G)|G^{\text{full}}) = E_{H_o'}(\exp(\delta w_i Z_i^G)|G^{\text{full}})$ Existing software such as GENEHUNTER (Kruglyak *et al.* 1996), GENEHUNTER-PLUS (Kong and Cox 1997), or ALLEGRO (Gudbjartsson *et al.* 2000) can be easily modified to make these calculations.

6.5 Assessment of significance conditional on a linkage result

In a linkage study for a complex trait, power to detect linkage to a given causal variant may not be high; some luck may often be involved in obtaining, say, a suggestive linkage result. ("*Suggestive linkage: statistical evidence that would be expected to occur one time at random in a genome scan. Significant linkage: statistical evidence that would be expected to occur 0.05 times in a genome wide scan.*" Lander and Kruglyak 1995). Suppose a particular polymorphism is the sole causal variant in the region, and suppose that the genetic model and study design are such that the power to detect linkage is low. Then in order to detect at least suggestive evidence for linkage, it may be necessary to have excess sharing even beyond what would ordinarily be expected under the genetic model for the causal variant. Suppose one later collects SNP data from the same individuals who were part of the linkage study, in a region showing linkage, and then applies our test. Then conditional on detection of at least suggestive linkage, there may be excess sharing that cannot be fully explained by the genotype data at the causal variant. Therefore, if one applies our test to only the data sets that have shown at least suggestive evidence for linkage, the test is no longer calibrated. For such cases, the significance of our test may need to be assessed conditional on the fact that suggestive evidence for linkage was exceeded.

Suppose there are n families in such a data set. Let $\mathbf{G} = (G_1, G_2, ..., G_n)$, where G_i is the genotype configuration for the affected individuals in the ith family. Let T be our test statistic (T_1, T_2 or T_3), and let W be the event that suggestive evidence for linkage was exceeded. The adjusted p-value of our test is then

$$P_{H_o}(T > t_{obs} | \mathbf{G}, W), \tag{6.3}$$

where H_o is the null hypothesis that the SNP is the sole cause in the region. One can assess (6.3) by simulation from $P_{H_o}(T|\mathbf{G}, W)$. For each replicate, conditional on the observed genotypes $\mathbf{G} = (G_1, G_2, ..., G_n)$, IBD sharing D_i by the i_{th} pair can be simulated based on $P_{H_o}(D_i|G_i, \text{both affected}) = P(D_i|G_i)$ given in Table 6.1. From this, linkage data for the rest of the region can be simulated. The linkage result and the test statistic T can be calculated, and the replicate is kept only if the linkage result exceeds suggestive evidence for linkage. The replicates that are not discarded are independent, identically-distributed draws from $P_{H_o}(T|\mathbf{G}, W)$, and the p-value given by (6.3) can then be estimated from this empirical distribution.

Chapter 7

Results and Discussion

7.1 Simulation studies

We perform simulation studies to assess the power of our method to detect that a given SNP is not the sole cause of linkage to the region. Each simulation involves 10^5 replicates of a data set of 150 affected sib pairs with complete IBD information. Simulations are performed under various genetic models, each of which involves epistasis among unlinked loci. In each case, we examine power to reject a non-causal locus that is completely linked ($\theta = 0$) to a causal locus, assuming various degrees of linkage disequilibrium ($D' = 0$, .5, or 1) and various allele frequencies. In each case, we use test statistic T_1 of equation (6.1), with $S = S_{pairs}$, $w = \sqrt{\sigma_G}$, and with significance assessed by a normal approximation. We also perform simulations to assess the adequacy (Type I error) of the normal approximation and find that it performs extremely well in these cases (results not shown). Specific details of the models follow.

To test the null hypothesis that a particular SNP is the sole cause in the region, we consider the exponential likelihood of Kong and Cox (1997). We consider three different disease models as described below. All the models allow epistasis.

Model I consists of two unlinked causal SNPs, both acting dominantly, with epistasis between them. In addition to the two allele frequencies, there are two penetrance parameters, p_1 and p_2 ($p_1 > p_2$), with penetrance p_1 for individuals who have both at least one copy of allele 1 at locus 1 and at least one copy of allele 1 at locus 2, and penetrance p_2 for all other individuals. (Penetrance is the probability that an individual is affected by the disease, given the observed genotype for that individual.) Model II consists of two unlinked causal SNPs, one (locus 1) acting recessively and the other (locus 2) following a general two-allele model, with epistasis between them. In addition to two allele frequencies, there are four penetrance parameters ($p_1 > p_2 > p_3 > p_4$), with penetrance p_1 for individuals who have genotype 1/1 at both locus 1 and locus 2, penetrance p_2 for those with both genotype 1/1 at locus 1 and genotype 1/2 at locus 2, penetrance p_3 for those with both genotype 1/1 at locus 1 and genotype 2/2 at locus 2, and penetrance p_4 for all other individuals. Model III consists of three unlinked causal SNPs, each acting dominantly, with epistasis among them. In addition to the three allele frequencies, there are two penetrance parameters ($p_1 > p_2$), with penetrance p_1 for individuals with both at least one copy of allele 1 at locus 1 and at least one copy of allele 1 at either locus 2 or locus 3, and with penetrance p_2 for all other individuals.

For any one of the three models above, with values chosen for the allele frequencies and penetrance parameters, we focus on causal locus 1, as defined above. We can obtain the joint distribution of $P((G^c, D^c) \mid \text{both affected})$, where G^c is the genotype configuration at causal locus 1, L_c, for an affected sib pair, and D^c is the number of alleles shared IBD by the pair at L_c. We first simulate 10^5 replicates of a data set of 150 affected sibs pairs from this distribution. Consider a non-causal SNP at locus L_n with genotype configuration G^n and IBD sharing D^n. We assume that L_n is completely linked ($\theta = 0$) with the causal locus L_c, so $D^n = D^c$. We then generate data G^n for the non-causal SNP at L_n, for the cases of linkage equilibrium ($D' = 0$), partial linkage disequilibrium ($D' = 0.5$) and complete linkage disequilibrium ($D' = 1$) with L_c, and for various choices of allele frequency. For each set of simulations, we test that the non-causal SNP at locus L_n is the sole cause of linkage to the region. The results are given in Table 7.1. It can be seen that in some cases, our method has substantial power to reject the null hypothesis for non-causal loci completely linked to a true causal locus, even when the non-causal locus is in complete linkage disequilibrium ($D' = 1$) with the causal locus. Note that if the non-causal locus is in perfect linkage disequilibrium with the causal locus (i.e. $|D'| = 1$ with the coupled alleles having identical allele frequencies), then the two loci cannot be distinguished based on the available data, and the probability of rejecting the null hypothesis for the non-causal loci is the same as the chosen Type I error.

7.2 Application to the $NIDDM1$ data set

Type 2 diabetes or non-insulin-dependent diabetes mellitus (NIDDM) is a complex disease affecting approximately 4% of the adult population worldwide. Linkage analysis and further fine mapping have localized a major susceptibility locus, denoted as $NIDDM1$, to a small region of chromosome 2. The $NIDDM1$ data set (Horikawa *et al.* 2000) is collected for the purpose of positional cloning of $NIDDM1$.

Model	f^c	f^n	D'	Haplotype frequency				Power at level	
				h_{11}	h_{12}	h_{21}	h_{22}	.05	.01
I	.2	.3	0	.06	.14	.24	.56	.9710	.8783
			.5	.13	.07	.17	.63	.8793	.6691
			.5	.03	.17	.27	.53	.9669	.8683
			1	.2	0	.1	.7	.2679	.0900
			1	0	.2	.3	.5	.9456	.8099
		.2	0	.04	.16	.16	.64	.9765	.8948
			.5	.12	.08	.08	.72	.7738	.5127
			.5	.02	.18	.18	.62	.9808	.9107
			1	.2	0	0	.8	.0495	.0101
			1	0	.2	.2	.6	.9810	.9146
II	.515	.3	0	.15	.36	.15	.34	.9811	.9165
			.5	.225	.29	.075	.41	.9718	.8859
			.5	.075	.440	.225	.260	.9381	.8019
			1	.3	.215	0	.485	.8450	.6257
			1	0	.515	.3	.185	.6365	.3622
		.515	0	.265	.25	.25	.235	.9788	.9084
			.5	.39	.125	.125	.36	.8604	.6421
			.5	.148	.367	.367	.118	.8874	.6939
			1	.515	0	0	.485	.0477	.0080
			1	.03	.485	.485	0	.1142	.0257
III	.271	.4	0	.11	.16	.29	.44	.8836	.6747
			.5	.19	.081	.21	.519	.7575	.4904
			.5	.055	.216	.345	.384	.8366	.5993
			1	.271	0	.129	.6	.2576	.0893
			1	0	.271	.4	.329	.6491	.3713
		.271	0	.073	.198	.198	.531	.8921	.6947
			.5	.172	.099	.099	.63	.6360	.3576
			.5	.037	.234	.234	.495	.8967	.7035
			1	.271	0	0	.729	.0493	.0096
			1	0	.271	.271	.458	.8671	.6552

Table 7.1: Power to detect a non-causal SNP

Power to detect that a SNP is not the sole cause of linkage to region, where the models are as described in the text with the causal SNP being locus 1 of the model in each case. f^c is the frequency of allele 1 at the causal SNP, f^n is the frequency of allele 1 at the completely-linked non-causal SNP, D' is disequilibrium between the two SNPs, and $h_{ij}, i, j = 1, 2$ is the population haplotype frequency, where i is the allele at the causal SNP and j is the allele at the non-causal SNP.

We analyze an *NIDDM1* data set that differs slightly from the data set of Horikawa *et al.* (2000) in that some additional markers are typed and some genotyping errors, apparent when markers are typed in all members of all families, are removed. The *NIDDM1* data set includes 170 sibships: 121 affected sib pairs, 34 affected sib trios, 12 affected sib quartets, 2 affected sib quintets, and 1 affected sib sextet. We consider 22 SNPs typed in a region of about $300,000$ DNA base pairs. Based on these 22 SNPs and 16 flanking microsatellites, the information on IBD sharing in the region is complete for most of the sibships.

When performing our test for a given SNP, we must cope with the fact that genotype data for some individuals may be missing at that SNP — that is, G may be incompletely observed even when complete information is available on D. In the case of an affected sib pair for which G is not completely observed for a particular SNP, we omit that pair from the analysis of that SNP. For sibships with ≥ 3 affected sibs, when G is not completely observed for a particular SNP, in most cases, we are able to reconstruct G from the observed genotype data at that SNP combined with the sharing information D in the region. For the remaining cases with ≥ 3 affected sibs, we impute the missing information on G in such a way that our test is guaranteed to be conservative (i.e. we impute the G that implies the highest level of IBD sharing, among those G consistent with the observed genotype data and D). Note, however, that conservativeness of this procedure is no longer guaranteed when the significance level is adjusted for detection of suggestive evidence for linkage.

For each of the 22 SNPs, to test the null hypothesis that the SNP considered is the sole cause of linkage to the region, we use test statistic T_2 of equation (6.2) with weights $w = \sqrt{\sigma_G}$. The p-value is assessed by simulation, using 10^7 replicates and assuming complete information on D and G, as described in Section 6.2. Using all the families, the p-value of the test for detection of linkage to the region is 1.78×10^{-5}, where linkage is detected using the exponential likelihood with weights $w = \sqrt{\sigma}$. However, when we consider each individual SNP, some families may be discarded because of missing genotype data, as described above, so the p-value for detection of linkage varies across the SNPs. To adjust the p-value of our test for detection of suggestive evidence for linkage, for each SNP, we first simulate 10^7 replicates of the non-missing families to determine the threshold value of the log–likelihood-ratio for suggestive evidence for linkage. The significance level for suggestive linkage is set to 7.4×10^{-4} (Lander and Kruglyak 1995). To obtain the conditional p-value of our test, we simulate until we obtain at least 10^4 realizations in which suggestive evidence for linkage is exceeded, and we calculate the conditional p-value as described in Section 6.4.

The results of the analysis are given in Table 7.2. The reported p-values for the test that a SNP is the sole cause of linkage to the region (last two columns of Table 7.2) are all one-sided, and $\hat{\delta} > 0$ is observed in all cases. Aside from two SNPs (SNP 66 and SNP 62), for which the sample size is small (≤ 125) because many individuals are untyped for those SNPs, all of the SNPs are rejected as being the sole cause of linkage, even after adjustment for suggestive evidence for linkage. SNPs 66 and 62 are rejected before adjustment but not after. Note that all of the SNPs are tightly linked to *NIDDM1*, and SNPs 22, 23, 25, 26, 29, and 38 all show significant linkage disequilibrium with *NIDDM1* (Horikawa *et al.*,2000). Thus, the information provided by our method is different from that provided by tests of linkage and linkage disequilibrium. Furthermore, this example illustrates that our test can reject a SNP as being the sole cause of linkage. Our analysis suggests that there may be more than one causal polymorphism in the region, or, alternatively, that the single causal polymorphism is not included in the set of SNPs typed.

In Horikawa *et al.* (2000), quite a number of the polymorphisms (16) examined show association with the disease. A smaller number of polymorphisms (6) show significant association with the evidence for linkage, as determined by the ability of genotypes at the polymorphism to partition the evidence for linkage. Functional studies subsequently confirm that at least one of these polymorphisms (UCSNP-43) encodes variation that affects expression of the CAPN10 protein (Baier *et al.* 2000; Yang *et al.* 2001). Our results here show that none of the individual polymorphisms studied are sufficient to account for the evidence for linkage. Thus, the information provided by our test here is different from that provided by the original tests proposed in Horikawa *et al.* (2000) which are essentially testing whether the observed variation is associated with evidence for linkage. For example, despite the fact that the evidence for linkage is entirely confined to UCSNP-43, we can conclusively reject the hypothesis that the segregation of the variation at UCSNP-43 can account, by itself, for the observed evidence for linkage. As noted above, our findings are not inconsistent with the hypothesis put forward in Horikawa *et al.* (2000), that combinations of variants at CAPN10 affect susceptibility to type 2 diabetes and generate the original evidence for linkage, but are also consistent with the possibility that untested variation elsewhere in the $NIDDM1$ region might fully account with the evidence for linkage. In that case, the causal variation is presumably in linkage disequilibrium with the region.

Map Order	Locus	f [a]	n [b]	Linkage p-value	p-value of our test	
					unadjusted	adjusted
1	SNP20	.85	153	3.57×10^{-5}	.0001337	.0394
2	SNP66	.88	124	5.95×10^{-5}	.0009932	.1048
3	SNP45	.94	163	1.58×10^{-5}	.0001234	.0285
4	SNP44	.94	164	2.32×10^{-5}	.0001009	.0376
5	SNP43	.73	160	2.01×10^{-5}	.0000001	.0004
6	SNP79	.97	161	2.66×10^{-5}	.0000244	.0247
7	SNP78	.94	162	2.03×10^{-5}	.0000558	.0291
8	SNP77	.92	161	1.58×10^{-5}	.0000522	.0228
9	SNP56	.57	149	4.40×10^{-5}	.0001638	.0157
10	SNP19	.56	161	1.47×10^{-5}	.0000347	.0042
11	SNP48	.55	154	1.64×10^{-5}	.0000303	.0033
12	SNP62	.81	125	6.27×10^{-5}	.0081385	.1174
13	SNP63	.76	130	3.50×10^{-5}	.0001566	.0197
14	SNP26	.92	162	2.04×10^{-5}	.0000356	.0137
15	SNP25	.50	156	4.07×10^{-5}	.0000322	.0054
16	SNP24	.98	162	1.92×10^{-5}	.0000053	.0201
17	SNP23	.85	158	1.67×10^{-5}	.0000556	.0084
18	SNP22	.61	158	1.56×10^{-5}	.0019207	.0253
19	SNP53	.90	155	6.80×10^{-5}	.0000026	.0161
20	SNP38	.62	154	5.62×10^{-5}	.0004898	.0196
21	SNP29	.77	151	1.48×10^{-5}	.0001107	.0074
22	SNP28	.56	156	0.46×10^{-5}	.0003044	.0057

Table 7.2: Results of the analysis of the $NIDDM1$ data set

[a] Frequency of the common allele. [b] The number of families used in the analysis. The linkage p-value is the p-value for the ordinary allele-sharing test of linkage applied to the non-missing families for that SNP, the unadjusted p-value for our test is the p-value for the test of H_O: the given SNP is the sole cause of linkage, and the adjusted p-value is conditional on detection of suggestive evidence for linkage at level 7.4×10^{-4}.

7.3 Discussion

We have developed a new statistical approach to guide positional cloning studies of qualitative traits. Assuming that many polymorphic sites have been identified and genotyped in a region showing strong linkage with a trait, we wish to determine which site (or combination of sites) in the region influences susceptibility to the trait. Our approach is to identify the polymorphisms whose genotypes could fully explain, in the statistical sense, the observed linkage to the region. We formulate a hypothesis test for which the null hypothesis is that a particular polymorphism is the sole cause of linkage to the region. By inverting this test, we construct a confidence set for the true causal site. The results of this approach provide information that is different from that provided by tests of linkage or association. Our method allows for a very general model for how the site influences the trait, including epistasis with unlinked loci, correlated environmental effects within families, and gene-environment interaction. Simulation studies show that the method can have high power to reject non-causal SNPs, even in cases when they are tightly linked and in complete linkage disequilibrium with the causal SNP. Application to an *NIDDM1* data set (Horikawa *et al.* 2000) lead to rejection of all SNPs in the set, suggesting that either there is more than one causal polymorphism in the region or else that the single causal polymorphism is not among those typed in the data set.

Our method of testing for a single causal SNP can be generalized, in principle, to any type of causal polymorphism (e.g. microsatellites) and to multiple tightly-linked causal loci. Details, for the case of affected sib pairs, can be found in Appendix E. However, note that the power of our method depends, in large part, on the values of

$E_A[S|G] - E_{H_O}[S|G]$ for the families in the study, where $E_A[\cdot|\cdot]$ denotes the conditional expectation calculated under the true genetic model. If G provides complete information on IBD sharing among the affecteds, then $E_{H_O}[S|G] = S = E_A[S|G]$, and the given family does not provide any information under our method. Similarly, when G provides close to complete information on S, power is low. This is more likely to occur when G is the genotype information on a single highly polymorphic locus or when G is the joint genotype information on several tightly-linked loci, than when G is the genotype information on a single SNP. Low power in such cases is the price paid for the lack of assumptions on the genetic model, in which we allow arbitrary mode of inheritance, epistasis with unlinked loci, correlated environmental effects within families, and gene-environment interaction. A method that would be powerful for highly polymorphic sites or combinations of sites could certainly be obtained with more assumptions on the genetic model.

7.4 Appendix E

In principle, our method can be generalized to any type of causal polymorphism (e.g., microsatellites) and to multiple tightly-linked causal loci. First consider a single polymorphic site with m alleles. The number of possible genotypes depends on m and the number of the affecteds. For a pair of relatives, there are $m + 4\binom{m}{2} + 6\binom{m}{3} + \binom{m}{4}$ possible genotypes which can be divided into 7 different categories as shown in Table 7.3. The conditional distribution of $P(D|G)$ for an affected sib pair is given in Table 7.3. Table 7.4 gives the null conditional mean and null conditional standard deviation of S, when S_{pairs} is used, for each of the 7 genotype categories.

G	D		
	0	1	2
($i\,i$ $i\,i$)	$\frac{f_i^2}{(1+f_i)^2}$	$\frac{2f_i}{(1+f_i)^2}$	$\frac{1}{(1+f_i)^2}$
($i\,i$ $i\,j$)	$\frac{f_i}{1+f_i}$	$\frac{1}{1+f_i}$	0
($i\,i$ $j\,j$)	1	0	0
($i\,j$ $i\,j$)	$\frac{2f_if_j}{1+f_i+f_j+2f_if_j}$	$\frac{f_i+f_j}{1+f_i+f_j+f_if_j}$	$\frac{1}{1+f_i+f_j+2f_if_j}$
($i\,i$ $j\,k$)	1	0	0
($i\,j$ $i\,k$)	$\frac{2f_i}{1+2f_i}$	$\frac{1}{1+2f_i}$	0
($i\,j$ $k\,l$)	1	0	0

Table 7.3: $P(D|G)$ for an affected sib pair with microsatellite data

The conditional distribution, $P(D|G)$, of the number of alleles shared IBD by an affected sib pair at a particular microsatellite, conditional on the sibs' genotype configuration G at that locus, where f_i is the frequency of allele i in the population.

G	μ_G	σ_G
$(\,i\;i \quad i\;i\,)$	$\frac{2}{1+f_i}$	$\frac{\sqrt{2f_i}}{1+f_i}$
$(\,i\;i \quad i\;j\,)$	$\frac{1}{1+f_i}$	$\frac{\sqrt{f_i}}{1+f_i}$
$(\,i\;i \quad j\;j\,)$	0	0
$(\,i\;j \quad i\;j\,)$	$\frac{2+f_i+f_j}{1+f_i+f_j+2f_if_j}$	$\frac{\sqrt{f_i+f_j+8f_if_j+2f_if_j(f_i+f_j)}}{1+f_i+f_j+2f_if_j}$
$(\,i\;i \quad j\;k\,)$	0	0
$(\,i\;j \quad i\;k\,)$	$\frac{1}{1+2f_i}$	$\frac{\sqrt{2f_i}}{1+2f_i}$
$(\,i\;j \quad k\;l\,)$	0	0

Table 7.4: μ_G and σ_G for an affected sib pair with microsatellite data

The null conditional mean, $\mu_G = E_{H_O}[S_{pairs}|G]$, and the null conditional standard deviation, $\sigma_G = \sqrt{Var_{H_O}(S_{pairs}|G)}$, of the sharing statistic S_{pairs} for an affected sib pair, given the sibs' genotype configuration G at a particular microsatellite, under the null hypothesis H_o that the microsatellite is the sole causal site in the region, where f_i is the frequency of allele i in the population.

To extend our method from a single causal locus to multiple tightly-linked causal loci in the region of interest, we assume that no crossovers occur within the sampled families, among the causal loci in the region. Under this assumption, the hypothesized causal loci would all have the same pattern of IBD sharing among the affecteds. To test the null hypothesis that a particular set of polymorphisms jointly explain the observed linkage to the region, a straightforward extension of our method would be as follows: let D be the IBD sharing among the affecteds in the region, and let $G = (G^1, ..., G^L)$ be the joint genotype data, where G^l is the genotype data for the affecteds at the l^{th} putative causal locus and L is the total number of hypothesized causal loci in the region. To obtain the conditional distribution of $P(D|G)$, one needs the marginal distribution of $P(D)$ and the conditional distribution of $P(G|D) = P(G^1, ..., G^L|D)$. To obtain the latter, one requires haplotype frequency estimates from an appropriate control population.

Bibliography

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

[2] Baier LJ, Permana PA, Yang X, Pratley RE, Hanson R, *et al*. (2000). A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. *J Clin Invest* 106:R69-R73.

[3] Baum LE (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* 3:1-8.

[4] Begleiter H, Reich T, Nurnberger J Jr, Li TK, Conneally PM, Edenberg H, Crowe R, *et al*. (1999). Description of the Genetic Analysis Workshop 11 Collaborative Study on the Genetics of Alcoholism. *Genet Epidemiol* 17 (Suppl).

[5] Blangero J, Göring HHH, Williams, JT, Dyer T, Almasy L (in press). Quantitative trait nucleotide analysis using Bayesian model selection. Genetic Analysis Workshop 12. *Genet Epidemiol (Suppl).*

[6] Boehnke M (1991). Allele frequency estimation from data on relatives. *Am J Hum Genet* 48:22-25.

[7] Boehnke M, Cox NJ (1997). Accurate inference of relationships in sib-pair linkage studies. *Am J Hum Genet* 61:423-429.

[8] Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998). Compre-

hensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* 63:861-869.

[9] Broman KW, Weber JL (1998). Estimation of pairwise relationships in the presence of genotyping errors. *Am J Hum Genet* 63:1563-1564.

[10] Cardon LR, Abecasis GR (2000). Some properties of a variance components model for fine-mapping quantitative trait loci. *Behav Genet* 30:235-243.

[11] Chakraborty R, Jin L (1993a). Determination of relatedness between individuals using DNA fingerprinting. *Hum Biol* 65:875-895.

[12] Cobbs G (1978). Renewal process approach to the theory of genetic linkage: case of no chromatid interference. *Genetics* 89:563-581.

[13] Denniston C (1975). Probability and genetic relationship: two loci. *Ann Hum Genet* 39:89-104

[14] Donnelly KP (1983). The probability that related individuals share some section of genome identical by descent. *Theor Pop Biol* 23:34-63.

[15] Ehm MG, Wagner M (1998). A test statistic to detect errors in sib-pair relationships. *Am J Hum Genet* 62:181-188.

[16] Epstein MP, Duren WL, Boehnke M (2000). Improved relationship inference for pairs of individuals. *Am J Hum Genet (Suppl)* 67:A308.

[17] Foss E, Lande R, Stahl FW, Steinberg CM (1993). Chiasma interference as a function of genetic distance. *Genetics* 133:681-691.

[18] Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999). Combined linkage and

association sib-pair analysis for quantitative traits. *Am J Hum Genet* 64:259-267.

[19] Göring HHH, Ott J (1997). Relationship estimation in affected sib pair analysis of late-onset diseases. *Eur J Hum Genet* 5:69-77.

[20] Greenberg DA (1993). Linkage analysis of "necessary" disease loci versus "susceptibility" loci. *Am J Hum Genet* 52:135-143.

[21] Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* 25:12-13.

[22] Hodge SE (1993). Linkage analysis versus association analysis: Distinguishing between two models that explain disease-marker associations. *Am J Hum Genet* 53:367-384.

[23] Horikawa Y, Oda N, Cox NJ, *et al.* (2000). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics* 26:163-172.

[24] Jacquard A (1974). The genetic structure of populations. Springer-Verlag, New York.

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

[26] Kong A, Cox NJ (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 61:1179-1188.

[27] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996). Parametric and non-

parametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 58:1347-1363.

[28] Lander ES, Green P (1987). Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 84:2363-2367.

[29] Lander ES, Kruglyak L (1995). Genetic dissection of complex traits: guildelines for interpreting and reporting linkage results. *Nature Genetics* 11:241-247.

[30] Lange K (1997). Mathematical and statistical methods for genetic analysis. Springer-Verlag, Berlin.

[31] Lin SL, Speed TP (1996). Incorporating crossover interference into pedigree analysis using the chi(2) model. *Hum Hered* 46:315-322.

[32] McPeek MS (1996). Introduction to recombination and linkage analysis. *IMA Volumes in Mathematics and its Applications*, Volume 81, Genetic mapping and DNA sequencing, Speed TP and Waterman MS, eds., Springer-Verlag.

[33] McPeek MS (in press). Inference on pedigree structure from genome screen data. *Statistica Sinica* (special bioinformatics issue).

[34] McPeek MS, Speed TP (1995). Modeling interference in genetic recombination. *Genetics* 139:1031-1044.

[35] McPeek MS, Sun L (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* 66:1076-1094.

[36] O'Connell JR, Weeks DE (1998). PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 63:259-266.

[37] Olson JM (1999). Relationship estimation by Markov-process models in a sib-pair linkage study. *Am J Hum Genet* 64:1464-1472.

[38] Ott J (1998). Analysis of human genetic linkage, revised edition. The Johns Hopkins University Press, Baltimore.

[39] Siegmund KD, Vora H, Gauderman WJ (in press). Combined linkage and linkage disequilibrium analysis in pedigrees. Genetic Analysis Workshop 12. *Genet Epidemiol (Suppl)*.

[40] Soria JM, Almasy L, Souto JC, *et al.* (2000). Linkage analysis demonstrates that the prothrombin G20210A mutation jointly influences plasma prothrombin levels and risk of thrombosis. *Blood* 95:2780-2785.

[41] Stam P (1979). Interference in genetic crossing over and chromosome mapping. *Genetics* 92:573-594.

[42] Sun L, Abney M, McPeek MS (in press). Detection of misspecified relationships in inbred and outbred pedigrees. Genetic Analysis Workshop 12. *Genet Epidemiol (Suppl)*.

[43] Sun L, Wilder K, McPeek MS (2001). Enhanced pedigree error detection. in preparation.

[44] Thompson EA (1974). Gene identities and multiple relationships. *Biometrics* 30:667-680.

[45] Thompson EA (1975). The estimation of pairwise relationships. *Ann Hum Genet* 39:173-188.

[46] Thompson EA (1986). Pedigree analysis in human genetics. The Johns Hopkins University Press, Baltimore.

[47] Thompson EA (1988). Two-locus and three-locus gene identity by descent in pedigrees. *IMA J Math Appl Med Biol* 5:261-280.

[48] Tiwari H, Elston R (1999). A new explicit algorithm to calculate identity by descent probabilities for pairs of relatives with respect to two linked loci. *Genet Epidemiol* 17:216.

[49] Whittemore AS (1996). Genome scanning for linkage: an overview. *Am J Hum Genet* 59:276-287.

[50] Yang X, Pratley RE, Baier LJ, Horikawa Y, Bell GI, Bogardus C, Permana PA (2001). Reduced skeletal muscle calpain-10 transcript level is due to a cumulative decrease in major isoforms. *Molec Genet and Metab* 73:111-113.

[51] Zhao H, Speed TP, McPeek MS (1995). Statistical analysis of crossover interference using the chi-square model. *Genetics* 139:1045-1056.