# STATISTICAL LINKAGE ANAYLSIS AND ASSOCIATION STUDIES

By

Martin Andrew Perry

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
SEPTEMBER 8 2000

© Copyright by Martin Andrew Perry, 2000

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

*Your file Votre référence*

*Our file Notre référence*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-57208-0

Canada

*I dedicate this to my parents,*

*John and Ann Perry,*

*without whose moral and financial support*

*this would not have been possible.*

# Contents

# List of Tables

# List of Figures

# Abstract

It is well known that many diseases have a genetic basis, or at least a genetic influence. One of the common problems in modern medicine is the determination of what diseases have a genetic influence and what part or parts of the human genome is implicated for a specific disease. This thesis investigates standard tests for association, as well as tests for linkage developed by Haseman and Elston (1972), Risch (1990) and several others. The transmission disequilibrium test (TDT) developed by Spielman et al (1993) to test for linkage disequilibrium as a result of linkage and association is also examined. The power of some of these tests is calculated and compared.

# Acknowledgements

I would like to thank my supervisor, Dr. David Hamilton, for giving me the motivation and assistance necessary for the completion of this thesis.

# Chapter 1

# Introduction

It is well known that many diseases have a genetic basis, or at least a genetic influence. One of the common problems in modern medicine is the determination of what diseases have a genetic influence and what part or parts of the human genome is implicated for that specific disease. This thesis investigates several tests for association and linkage between a marker locus and a disease locus.

## 1.1 Background

Some diseases in humans have an obvious genetic component. Colour blindness and hemophilia have long been recognized to be hereditary (and thus are genetically determined). Other diseases such as cystic fibrosis and Down's syndrome have also been solely attributed to genetic causes.

There are a lot of other diseases that are believed to have a significant genetic component that do not lend themselves to simple genetic analysis. Diseases such as osteoporosis, heart disease, alcoholism, hyperactivity disorder and bipolar mood disorder are only a few of the disease that are believed to have a genetic component to them.

Finding what part of the genome influences these diseases can be a complicated task. The diseases may be influenced by several different parts of the genome and

in some cases people with identical genomes (identical twins) may have a different disease status (one affected, one not). This is possible because most of these diseases are also known to depend on environmental factors. A simple example of this is that if a person never consumes alcohol, they could never become an alcoholic. Also, a person who exercises and eats well may avoid osteoporosis and heart disease.

In order to determine what part of the human genome influences diseases like these, it is necessary to use statistical methods.

## 1.2 Basic Genetics

The genetic information of all organisms are contained in the deoxyribonucleic acid (**DNA**) of that organism. The "recipe for life" is contained in the long strings of DNA that are referred to as **chromosomes**. DNA has a, now famous, double helix structure and consists of four "bases", adenine, cytosine, guanine and thymine, which are most often referred to by A, C, G and T. Different sections of the chromosomes have different purposes. **Genes** are DNA sequences that lead to the production of particular products (such as proteins), and a gene can be thought of as a discrete unit of information influencing inherited characteristics. The chromosomal location of a gene is referred to as a **locus** (pl. loci) (Lynch and Walsh 1998).

Most organisms have two copies of each chromosome. These organisms are said to be **diploid**. Organisms or cells that have only one copy of each chromosome are said to be **haploid**. Humans are diploid organisms that have 23 pairs of chromosomes, one pair of sex chromosomes and 22 pairs of **autosomal** (not sex related) chromosomes. The sex chromosomes are commonly referred to as X and Y. Females have two X chromosomes, while males have both an X and a Y chromosome. Since colour blindness and hemophilia affect males and females in significantly different proportions, but are known to be hereditary, they are influenced by genes on the sex chromosome.

For sexual reproduction to occur the diploid cells must prepare for sexual reproduction in a process referred to as **meiosis**. During meiosis, the diploid cells that will

be involved in reproduction (gametes) divide into two individual cells, each containing one of the two chromosomes in the parent. For normal fertilization to occur, the haploid cell of the male (sperm) must join with the haploid cell of the female (egg) to form a new diploid cell. This new diploid cell may then grow into a new individual. Thus, one half of the genetic information of each individual comes from each parent. Down's syndrome is now known to be the result of a flaw in the meiotic process in which the affected individual has three of a particular chromosome (trisomy of the 21st chromosome)(Khoury, Beaty, and Cohen 1993).

It is only relatively recently that we have been able to examine the genetic material of individuals. This led to the discovery of **alleles** which are detectable variations occurring at a particular genetic locus. Since it is possible to detect what alleles an individual has at a specific locus, alleles are what are commonly used to describe the genetic make-up of an individual. Although much of the DNA in humans is thought to serve no purpose, there are an estimated 50,000-100,000 genes within the human genome, according to the National Human Genome Research Institute (http://www.nhgri.nih.gov/HGP/).

Statistical genetics is generally concerned with loci that have more than one allele. Loci that exhibit more than one allele are said to be **polymorpic**, whereas loci that only have one allele are **monomorpic**. Due to the fact that all the alleles are identical at a monomorphic loci, there is no information to be gained from that locus. Fortunately, a substantial fraction of loci are polymorphic to some degree. New or different alleles at a specific locus can result from a **mutation** (a change in the base pairs). Cystic fibrosis is a recessive disease that results from a mutation at a single locus. A **recessive** disease requires both alleles at the disease locus to be "defective". This is compared to a **dominant** disease, which requires only one allele to be "defective". Autosomal dominant polycystic kidney disease (ADPKD), which leads to the formation of cysts in the kidneys that can result in renal failure and/or death, is an example of a dominant disease.

The **genotype** of an individual is the particular set of alleles an individual has,

either at a specific locus or in reference to several or all loci. When a diploid organisms has the same allele on both chromosomes, that organism is said to be **homozygous** at that locus. Individuals with differing alleles at a locus are said to be **heterozygous** at that locus. For example, if we consider a locus with two alleles A and B, then there are three possible genotypes at that locus. These are the two homozygotes: AA and BB; and the heterozygote AB.

The simplest tests for determining what parts of the genome affect a specific disease are known as tests for association.

## 1.3 Test for Disease-Allele Association

Detecting an association between a marker locus and a disease can be a critical first step towards the identification of the genetic basis for the disease in question. **Disease-allele association** occurs when an allele occurs more frequently or less frequently in individuals affected with the disease than in unaffected individuals. There can be a positive association, in which the allele occurs more frequently in individuals who are affected, or a negative association, in which the allele shows up less frequently in affected individuals(Lynch and Walsh 1998).

Association is a lack of independence, in which the event of having the disease is not independent of the event of having the allele under consideration. The simplest test for association is thus a test for homogeneity in a contingency table.

Because many diseases of interest are quite rare, a case-control study is usually used. To do this, a random sample of affected individuals (cases) and a random sample of unaffected individuals (controls) are genotyped at the marker locus and the number of each type of allele is recorded. If we are only interested in a particular allele at the marker locus, allele $M_1$, then we can simplify matters by using $M_2$ to refer to all other alleles at that locus.

Let $n_A$ be the number of affected individuals (cases) and $n_U$ be the number of unaffected individuals (controls). Furthermore, let $n_{1A}$ be the number of times $M_1$

appears in the affected individuals and let $n_{2A}$ be the number of times $M_2$ appears at the marker locus of the affected individuals. Since we are dealing with the number of times the allele is present, $n_{1A} + n_{2A} = 2n_A$, because each individual will have two alleles at the marker locus. The alleles on each chromosome of an individual are assumed to be statistically independent because of random mating in the parents. The data is displayed in Table 1.3.

| | Cases | Controls | |
|---|---|---|---|
| Number of $M_1$ Alleles | $n_{1A}$ | $n_{1U}$ | $n_1$ |
| Number of $M_2$ Alleles | $n_{2A}$ | $n_{2U}$ | $n_2$ |
| | $2n_A$ | $2n_U$ | N |

Table 1.1: Counts for Simple Association Test

We could also test more than one allele at once. To test R alleles, we could set up an $R \times 2$ table with the counts of alleles on the rows of the table. The null hypothesis is that there is no association between the marker allele(s) and the disease. The null and alternative hypotheses are

$H_O$: Distributions are equal ($p_{Ai} = p_{Ui}$ for all $i$)
$H_A$: Distributions are not equal ($p_{Aj} \neq p_{Uj}$ for some $j$)

For the 2 × 2 case, the p-value can be calculated exactly using a hypergeometric distribution or approximated using a normal or $\chi^2$ distribution. Similarly for tables of greater dimension, Fisher's exact test can be used, or the goodness of fit statistic can be compared to its approximating distribution ($\chi^2$ with R-1 degrees of freedom).

Although association may be indicative of some biological relation between the allele and the disease (or at least the disease susceptibility locus), this simple test has limited application. This is because association can result for several other reasons.

### 1.3.1 Example of a Test for Association

As and example of a test for association, we use data that was presented as an example by at the June 2000 meeting of the Summer Institute of Statistical Genetics at the Duke Center for Human Genetics.

The data was given as an example of trying to find an association between the apoE4 (apolipoprotien E4) allele and late onset Alzheimer's disease in Caucasians. It should be noted that since the data was presented as an example, this may not be actually observed counts.

The data was

| **Oberved Counts** | Cases | Controls | Total |
|---|---|---|---|
| apoE4 | 240 | 60 | 300 |
| Not apoE4 | 360 | 340 | 700 |
| Total | 600 | 400 | 1000 |

Table 1.2: Observed Counts for Alzheimer's Example

| **Expected Counts** | Cases | Controls | Total |
|---|---|---|---|
| apoE4 | 180 | 120 | 300 |
| Not apoE4 | 420 | 280 | 700 |
| Total | 600 | 400 | 1000 |

Table 1.3: Expected Counts for Alzheimer's Example

The goodness of fit statistic is

$$\begin{aligned} T &= \frac{(240-180)^2}{180} + \frac{(60-120)^2}{120} + \frac{(360-420)^2}{420} + \frac{(340-280)^2}{280} \\ &= 71.61. \end{aligned}$$

The P-value is

$$P(\chi^2 \geq 71.61) \approx 10^{-16}.$$

Thus, there is overwhelming evidence (in this example) that allele apoE4 is associated with late-onset Alzheimer's disease.

## 1.4 Causes of Association

Association can be the result of a biological involvement of the allele with the disease. This can occur in several ways, the simplest of which is that the allele itself has an influence on the disease. If the allele causes, prevents, or influences the disease directly, then the allele will be associated with the disease, although if the effect of the allele is very small, the association may be too small to detect.

Although biological involvement of the allele directly may occur, it is more likely that this is not the case. What is more often the case is that the locus of the marker allele is close to the disease-influencing locus.

Association can also result from several different phenomenon. These are usually divided into several groups: random genetic drift, admixture, mutation, and the founder effect(Khoury, Beaty, and Cohen 1993). These divisions are by no means exclusive, as we show below.

**Genetic drift** is defined as cumulative changes in gene frequency due to sampling variation (Khoury, Beaty, and Cohen 1993). Random genetic drift is the situation in which, although the allele is not related to the disease, or close to the disease locus, the allele happens (by random chance) to occur more or less frequently in the affected individuals.

**Admixture** is the joining of several subpopulations into a single population. When multiple subpopulations are mixed, we can get disease-allele associations when the allele is not related to the disease in any relevant biological way. Consider the situation in which we have two subpopulations. If one of the populations has a greater risk of the disease in question, then any allele that is more or less common in that subpopulation than the other may be statistically associated with the disease (Lynch and Walsh 1998).

A mutation can lead to a spurious association as well. Consider a population in which a new allele is created in a single ancestral chromosome by mutation. Since mutation into a specific new allele at a particular locus is an extremely rare event, the only time that allele appears will be in individuals who have as a common ancestor

the individual in whom the mutation originally took place.

If the person who originally had the mutation is a member of a high-risk (or low-risk) population for the disease, although the mutation may have nothing to do with the disease it would possibly show up more (or less) often in affected individuals than in the unaffected individuals. Thus, although the allele has no affect on the disease there is an association between the allele and the disease. This is an extreme form of admixture.

Another related phenomenon that can lead to an allele-disease association is the **founder effect**. The founder effect occurs when a subpopulation is reduced to a small number before the subpopulation increases in size, for example when a small group of individuals migrates to an isolated geographic region. This can lead to chance associations in the expanding population as a result of genetic drift, because when the population is reduced to a small number, the number of alleles in the population is reduced. When the population increases in size, there are still only a relatively small number of alleles, so there may be very little genetic variation in the population. If there exists in the founders some alleles that predispose the individuals to the disease in question, there may be a high risk of the disease in the population.

The term founder effect is usually used to refer to subpopulations that are not breeding with other groups, such as the Amish, whereas the term admixture is generally used in interbreeding populations. Ellis van Creveld syndrome is an example of a disorder that has been attributed to a founder effect, as it is much more common in the Old Order Amish than in the general population (Khoury, Beaty, and Cohen 1993). Any test for association in the general population may find association because the disease is much more common in the Old Order Amish. Unless extreme care was taken, the cases would consist of mainly Amish, whereas the controls would come mainly from non-Amish individuals who may have a different allele frequency than the Amish.

Mutation, admixture and the founder effect are all forms of population stratification. **Population stratification** occurs when the total population consists of several

subpopulations which differ in both candidate gene (marker allele) frequency and risk of disease occurrence (Khoury, Beaty, and Cohen 1993). Most of the problems of finding an association that is not caused by biological involvement can be avoided by careful selection of the cases and controls. If the controls come from a different "population" than the controls (e.g. different ethnic origin or a closed breeding population such as the Amish) then the marker allele frequencies may be different in the two populations which could lead to a spurious association. Of course, finding an association in a subpopulation does not allow you to make inferences about the effect of that allele in the general population.

## 1.5 An Example of a Spurious Association

Consider two subpopulations A and B that form a population. For notational purposes, M (and m) will denote the presence (and absence) of the marker allele of interest and D (and d) will denote the presence (and absence) of the disease.

Suppose that subpopulation A makes up 90% of the population, $P(A) = 0.9$, and that $\mathrm{P(M}|A) = 0.8$, which says that 80% of the marker alleles in population A are the M allele. Also suppose that the proportion of people in subpopulation A who get the disease is 0.1 ($\mathrm{P(D}|A) = 0.1$). For subpopulation B, which makes up 10% of the population, let $\mathrm{P(M}|B) = 0.4$ and $\mathrm{P(D}|B) = 0.3$. Although the differences in these proportions may seem extreme, they can be used to demonstrate the problem of population stratification. Even with the assumption that there is no disease-allele association in either of the subpopulations, we can show that $P(M|D) \neq P(M|d)$, which is the alternative hypothesis for the test, so there will be association in the population.

The conditional probability for the marker given the disease is

$$P(M|D) = \frac{P(M \cap D)}{P(D)}$$

$$\begin{aligned} &= \frac{P(M \cap D \cap A)}{P(D)} + \frac{P(M \cap D \cap B)}{P(D)} \\ &= \frac{P(M \cap D|A)P(A)}{P(D)} + \frac{P(M \cap D|B)P(B)}{P(D)}. \end{aligned}$$

But M and D are conditionally independent given the subpopulation, so

$$\begin{aligned} P(M|D) &= P(M|A)\frac{P(D|A)P(A)}{P(D)} + P(M|B)\frac{P(D|B)P(B)}{P(D)} \\ &= P(M|A)P(A|D) + P(M|B)P(B|D). \end{aligned}$$

Substituting numerical values gives

$$P(D) = P(D|A)P(A) + P(D|B)P(B) = 0.12,$$

so

$$P(B|D) = \frac{P(BD)}{P(D)} = \frac{P(D|B)P(B)}{P(D)} = \frac{(0.3)(0.1)}{0.12} = 0.25.$$

Thus P(A|$D$) = 0.75, and

$$\begin{aligned} P(M|D) &= P(M|A)P(A|D) + P(M|B)P(B|D) \\ &= (0.8)(0.75) + (0.4)(0.75) \\ &= 0.7. \end{aligned}$$

Similarly P(M|d) can be expressed as

$$\begin{aligned} P(M|d) &= \frac{P(M \cap d \cap A)}{P(d)} + \frac{P(M \cap d \cap B)}{P(d)} \\ &= \frac{P(M \cap d|A)P(A)}{P(d)} + \frac{P(M \cap d|B)P(B)}{P(d)}. \end{aligned}$$

As before, we use the conditional independence of the marker and the disease given the subpopulation to show

$$\begin{aligned} P(M|d) &= \frac{P(M \cap d|A)P(A)}{P(d)} + \frac{P(M \cap d|B)P(B)}{P(d)} \\ &= P(M|A)\frac{P(d|A)P(A)}{P(d)} + P(M|B)\frac{P(d|B)P(B)}{P(d)} \\ &= P(M|A)P(A|d) + P(M|B)P(B|d). \end{aligned}$$

Substituting numerical values gives

$$P(B|d) = \frac{P(Bd)}{P(d)} = \frac{P(d|B)P(B)}{1 - P(D)} = \frac{(1 - 0.3)(0.1)}{1 - 0.12} = 0.0795.$$

Thus,

$$\begin{aligned} P(M|d) &= P(M|A)P(A|d) + P(M|B)P(M|d) \\ &= (0.8)(1 - 0.0795) + (0.4)(0.0795) \\ &= 0.7682. \end{aligned}$$

This shows that although the marker does not affect the disease status in any direct way, we could still detect an association if there was population stratification. It is comforting to know, however, that the problems associated may be removed after several generations of random mating. This is because the separate subpopulations will "blend" into a single population.

Consider the previous populations A and B in a truly random mating situation (no preference for mates based on previous subpopulation). We could think of the offspring of the matings as forming three groups based on which subpopulation their parents were in. If both the parents of the offspring are from subpopulation A or B, then we denote the new groups AA or BB, respectively. However, if one parent is from each population, consider that offspring to be part of a new group, AB.

Offspring will be in group AA with probability $P(A)^2$ (thus $P(AA) = (0.9)^2 = 0.81$) and will be in group BB with probability $P(B)^2 = 0.01$. Following this, $P(AB) = 2P(A)P(B) = 0.18$. In order to determine the effect of the random mating, we must know the probability of getting the disease for these three groups. In the

case where both parents are from the same subpopulation, it makes sense to say that the probability of getting the disease is the same as it was in the parents' subpopulation. Thus $P(D|AA) = P(D|A) = 0.1$ and $P(D|BB) = P(D|B) = 0.3$. For the new group AB, we have to arbitrarily set $P(D|AB)$ for the purpose of this example. The actual risk of disease for people in group AB would depend on the mode of inheritance (ie. dominance/recessive, number of loci involved). It is reasonable to assume that $P(D|AB)$ would be between $P(D|A)$ and $P(D|B)$, so we will let $P(D|AB) = 0.2$.

In order to see the effect random mating has on the association in the population, we need to calculate both P(M|D) and P(M|d) for the offspring.

$$\begin{aligned} P(M|D) &= \frac{P(M \cap D \cap AA)}{P(D)} + \frac{P(M \cap D \cap Ab)}{P(D)} + \frac{P(M \cap D \cap BB)}{P(D)} \\ &= \frac{P(M \cap D|AA)P(AA)}{P(D)} + \frac{P(M \cap D|AB)P(AB)}{P(D)} + \frac{P(M \cap D|BB)P(BB)}{P(D)}. \end{aligned}$$

Once again, because of the conditional independence of marker and disease given the group,

$$P(M|D) = P(M|AA)P(AA|D) + P(M|AB)P(AB|D) + P(M|BB)P(BB|D).$$

Now we can calculate the overall disease prevalence

$$\begin{aligned} P(D) &= P(D|AA)P(AA) + P(D|AB)P(AB) + P(D|BB)P(BB) \\ &= (0.1)(0.81) + (0.2)(0.18) + (0.3)(0.01) = 0.12. \end{aligned}$$

It is interesting to note that because we chose $P(D|AB) = 0.2$, the average between the two subpopulations, the proportion of people in the total population who get the disease did not change. Also, the conditional probabilities of the subpopulations given the disease are

$$P(AA|D) = \frac{P(AA \cap D)}{P(D)} = \frac{P(D|AA)P(AA)}{P(D)} = \frac{(0.1)(0.81)}{0.12} = 0.675$$

$$
\begin{aligned}
P(BB|D) &= \frac{P(BB \cap D)}{P(D)} = \frac{P(D|BB)P(BB)}{P(D)} = \frac{(0.3)(0.01)}{0.12} = 0.025 \\
P(AB|D) &= 1 - P(AA|D) - P(BB|D) = 0.3.
\end{aligned}
$$

Since each parent contributes one half the genes to each offspring, $P(M|AB) = P(M|A)/2 + P(M|B)/2 = 0.6$. We can now evaluate

$$
\begin{aligned}
P(M|D) &= P(M|AA)P(AA|D) + P(M|AB)|P(AB|D) + P(M|BB)|P(BB|D) \\
&= (0.8)(0.675) + (0.6)(0.3) + (0.4)(0.0275) \\
&= 0.731.
\end{aligned}
$$

As before, we want to compare this to $P(M|d)$. Using simplifications like before, we can write

$$
\begin{aligned}
P(M|d) &= P(M|AA \cap d)P(AA|d) + P(M|AB \cap d)|P(AB|d) + P(M|BB \cap d)|P(BB|d) \\
&= P(M|AA)P(AA|d) + P(M|AB)|P(AB|d) + P(M|BB)|P(BB|d).
\end{aligned}
$$

Also, the conditional probabilities of the subpopulations given the disease is not present are

$$
\begin{aligned}
P(AA|d) &= \frac{P(AA \cap d)}{P(d)} = \frac{P(d|AA)P(AA)}{1 - P(D)} = \frac{(1 - 0.1)(0.81)}{1 - 0.12} = 0.828 \\
P(BB|d) &= \frac{P(BB \cap d)}{P(d)} = \frac{P(d|BB)P(BB)}{1 - P(D)} = \frac{(1 - 0.3)(0.01)}{1 - 0.12} = 0.008 \\
P(AB|d) &= 1 - P(AA|d) - P(BB|d) = 0.164.
\end{aligned}
$$

So the desired probability is

$$
\begin{aligned}
P(M|d) &= P(M|AA)P(AA|d) + P(M|AB)|P(AB|d) + P(M|BB)|P(BB|d) \\
&= (0.8)(0.828) + (0.6)(0.164) + (0.4)(0.008) \\
&= 0.764.
\end{aligned}
$$

We can see that before there was the random mating the difference in probabilities of the marker allele between the diseased and nondiseased groups was $P(M|d) - P(M|D) = 0.7682 - 0.7 = 0.0682$. After one generation of random mating, the difference is reduced to $0.764 - 0.731 = 0.033$. The difference is approximately halved in one generation, with both $P(M|d)$ and $P(M|D)$ approaching the population proportion, $P(M) = P(M|A)P(A) + P(M|B)P(B) = 0.76$. After several generations of random mating, the association would be undetectable unless extreme measures were taken to detect it. The reduction in the association depends on the probability of disease in the new group. Most human populations however do not practice strictly random mating.

# Chapter 2

# Linkage Analysis

Linkage analysis is used to determine if a marker locus is linked to a disease-influencing locus. Two loci are said to be linked of they are on the same chromosome and close together. If two loci are linked, then the transmission of alleles from parent to offspring at the loci are not independent of each other. In tests of linkage, the most common subjects are affected sib pairs (ASPs), as they were the first type of subjects used to study linkage and there exists a relatively straightforward test statistic (Lynch and Walsh 1998).

Affected sib pairs are a pair of siblings (brother/brother, brother/sister or sister/sister) who are both affected by the disease. The basic idea of the test is that, if there is a genetic influence to the disease, the two affected sibs probably have the same alleles at the disease locus. If a marker locus is linked to the disease locus, then the ASPs probably have the same alleles at the marker locus as well. The test compares the number of alleles that are the same at the marker locus to the number of alleles that are expected to be the same if there is no linkage.

Affected sib pair tests also use the idea that the transmission of alleles from parents to offspring follows known patterns based on simple Mendelian genetics. Using Mendelian genetics, we can calculate the probability of the siblings having alleles that are **identical by descent (IBD)**. Two alleles are IBD if they both descended from a common ancestor. In the case of sib pairs, two alleles in the offspring are IBD if

they both came from the same chromosome of the same parent.

Since there are two alleles at each locus, siblings can have 0, 1 or 2 alleles IBD and the number they have IBD is called the IBD status of the sib pair. The probability that a sib pair has a given IBD status can be calculated by considering the offspring of two heterozygous parents with alleles $M_1M_2$ and $M_3M_4$ at a locus. Ignoring disease status, all possible pairs of offspring can be enumerated and the IBD status determined, as shown in Table 2.

| | | First Sibling | | | |
|---|---|---|---|---|---|
| | | $M_1M_3$ | $M_1M_4$ | $M_2M_3$ | $M_2M_4$ |
| Second Sibling | $M_1M_3$ | 2 | 1 | 1 | 0 |
| | $M_1M_4$ | 1 | 2 | 0 | 1 |
| | $M_2M_3$ | 1 | 0 | 2 | 1 |
| | $M_2M_4$ | 0 | 1 | 1 | 2 |

Table 2.1: IBD Status for Siblings of Heterozygous Parents

Each combination is equally likely under the standard biological assumptions of equal transmission probabilities of alleles (each allele is assumed to have a probability of being passed on of 1/2), so it is possible to calculate the probability of the IBD status. Four of the sixteen combinations have no alleles IBD, so $P(IBD = 0) = 4/16 = 1/4$. Similarly $P(IBD = 1) = 8/16 = 1/2$ and $P(IBD = 2) = 4/16 = 1/4$.

There are other ways in which these probabilities can be calculated, the easiest of which is to consider the paternal and maternal alleles of the offspring separately. Under the standard assumption of Mendelian genetics, the maternal alleles and paternal alleles are passed independently to each offspring. The probability that the same paternal allele gets passed to both offspring is 1/2, and this is the same as the probability that the same maternal allele gets passed to both offspring. Since these events are independent, with the same probability, $p = 1/2$, and there are a fixed number of trials, $n = 2$, a binomial distribution can be used to obtain the probabilities.

Either of the above approaches can be used to calculate IBD status probabilities

for other types of relatives as well. The probabilities are presented in Table 2.2.

| Type of Relative Pair | Probability of IBD Status | | |
|---|---|---|---|
| | 0 | 1 | 2 |
| Monozygotic (Identical) Twins | 0 | 0 | 1 |
| Full Sibs | 1/4 | 1/2 | 1/4 |
| Parent-Offspring | 0 | 1 | 0 |
| First Cousins | 3/4 | 1/4 | 0 |
| Double First Cousins | 13/16 | 1/8 | 1/16 |
| Grandparent-Grandchild | 1/2 | 1/2 | 0 |
| Aunt/Uncle-Nephew/Niece | 1/2 | 1/2 | 0 |

Table 2.2: Probabilities of IBD status for Different Relative Pairs (Khoury, Beaty, and Cohen 1993)

Double first cousins arise when a sibling pair breeds with another sibling pair. For example, when two brothers marry two sisters, the offspring in the two families will be first cousins on both the maternal and paternal side, hence double first cousins.

In order to determine the IBD status of a pair of offspring, it is necessary to determine the genotype of the sibs, as well as the parents at the marker locus. However, it is often not possible to unambiguously determine the IBD status of a pair of siblings. If one of the parents is a homozygote at the marker locus, then it cannot determined whether the siblings are IBD for the alleles that descended from that parent because we cannot tell if the alleles came from the same chromosome. Other situations exist in which we would not be able to determine the IBD status for a pair of offspring as well. Consider the situation in which the parents are both heterozygous at the marker locus, but they both have the same alleles, A and B. The offspring of these parents could have three genotypes at the marker locus: AA, AB and BB. Suppose these parents have two heterozygous offspring (both AB). It is not possible to unambiguously determine the IBD status for this sib pair, as we can not distinguish which allele came from which parent. However, if both offspring are homozygotes then the

IBD status could be determined, which would be 2 if the offspring are homozygous for the same allele and 0 if they are not.

Determining the IBD status of more distant relatives is even more problematic, as many individuals must be genotyped and it is often more difficult to obtain an informative pair. For example, in a grandparent-grandchild relation, we would have to type not only the grandparent and grandchild, but both parents of the grandchild and the other grandparent (on the same side of the family) as well. To be able to determine if they have an allele IBD, we would have to determine what allele the grandparent passed to the parent, and we would have to be able to determine if that allele was passed to the grandchild.

The lack of availability of individuals for testing creates problems for many diseases, particularly those with late onset. Osteoporosis and Alzheimer's disease are good examples of late-onset disease that are believed to have a genetic component. As mentioned earlier it is necessary to know what alleles both parents had at the marker locus, to determine the IBD status of affected siblings. The parents, however, are not likely to be available for genotyping as one or both may have died.

Not being able to determine IBD status creates a serious problem for most simple tests for linkage. Only informative affected relative pairs can be used, which are relative pairs for which the IBD status can be determined. If a marker locus is highly polymorphic (many alleles), then the probability that a relative pair is informative increases. This is because the individuals involved are both less likely to be homozygotes, and less likely to have the same alleles, both of which can create problems in determining IBD status. Thus, when doing linkage studies it is beneficial to use a polymorphic marker locus as this will make getting a sufficient sample size easier.

## 2.1 Some Common Tests for Linkage with Affected Relative Pairs

There are many different tests for linkage with affected relative pairs in use, and for the purposes of this thesis we will mainly focus on two common tests. All tests, however, test the hypotheses

$H_O$: No linkage between the marker and disease locus.
$H_A$: There is linkage between the marker and disease locus.

The data consists of the three counts, $n_0$, $n_1$ and $n_2$, corresponding to the number of affected relative pairs with 0, 1 or 2 alleles IBD respectively. The simplest test for linkage in ASPs only uses $n_2$, the number of ASPs who have two alleles IBD at a marker locus (Lynch and Walsh 1998). The hypotheses can be restated as

$H_O : p_2 = \frac{1}{4}$
$H_A : p_2 > \frac{1}{4}$

where $p_2$ is the probability of an ASP having an IBD status of 2.

Under $H_O$, $n_2$ is binomial $(n, 1/4)$, where $n = n_0 + n_1 + n_2$ is the total number of affected pairs. An exact P-value can be calculated as

$$P = P(n_2 \geq n_{2,obs}).$$

For large $n$, $n_2$ is approximately normal with mean $n/4$ and variance $3n/16$, so the statistic

$$T_1 = \frac{n_2 - n/4}{\sqrt{3n/16}} \tag{2.1}$$

is approximately standard normal. The approximate P-value is

$$P = P(T_1 \geq T_{1obs}).$$

It is a one sided test, due to the fact that there is linkage between the marker locus and the disease locus only if there is an excess of sib pairs with 2 alleles IBD at the marker locus. Similar tests can be done for any type of relative pair, although for most types of relative pairs it is impossible to have more than one allele IBD.

Another test statistic for affected relative pairs, proposed by Risch (1990b), uses the LOD score (likelihood of odds). The LOD score is simply the common logarithm of the likelihood ratio statistic for the hypotheses

$H_O : p = \alpha_{R0}$
$H_A : p < \alpha_{R0}$

where the probability $\alpha_{R0}$ is obtained from the first column of Table 2.2. The test statistic focuses on the number of relative pairs who have 0 alleles IBD, mostly because it is not possible for relative pairs other than sibs to have two alleles IBD. The maximum LOD score statistic (MLOD) is

$$T_{MLOD} = n_0 \log_{10}\left(\frac{n_0}{n\alpha_{R0}}\right) + (n - n_0)\log_{10}\left(\frac{n - n_0}{n - n\alpha_{R0}}\right) \tag{2.2}$$

which is 2.303 times smaller than the usual log likelihood ratio. If $T_{MLOD} > 3$ then the evidence for linkage is assumed to be significant, which is a standard practice in the genetics literature for LOD scores. This is a more stringent requirement than an $\alpha = 0.05$ significance level. Using the usual $\chi^2$ approximation for twice the negative of the natural logarithm of the likelihood ratio shows that the significance level is approximately 0.0001.

Although these tests are relatively simple to use, calculating their power can be quite difficult.

### 2.1.1 An Example of Linkage Tests Using Affected Sib Pair

For an example of these linkage tests, consider the data from the study of Walker and Cudworth (1980) that was given in Motro and Thomson (1985). The study used

119 affected informative sib pairs to try to find linkage between a marker locus and IDDM (Insulin Dependent Diabetes Mellitus).

At the marker locus, the data was $n_2 = 69$, $n_1 = 43$ and $n_0 = 7$.

Using test statistic (2.1) first,

$$\begin{aligned} T_{1obs} &= \frac{n_2 - n/4}{\sqrt{3n/16}} \\ &= \frac{69 - 119/4}{\sqrt{357/16}} \\ &= \frac{39.25}{4.72} = 8.31 \end{aligned}$$

Thus the approximate P-value is

$$P = P(T_1 \geq 8.31) \approx 0$$

which is significant.

Using The MLOD test (2.2), gives

$$\begin{aligned} T_{MLODobs} &= n_0 \log_{10}\left(\frac{n_0}{n\alpha_{R0}}\right) + (n - n_O)\log_{10}\left(\frac{n - n_0}{n - n\alpha_{R0}}\right) \\ &= 7\log_{10}\left(\frac{7}{119/4}\right) + (119 - 7)\log_{10}\left(\frac{119 - 7}{119 - 119/4}\right) \\ &= -4.39877 + 11.04422 = 6.64545. \end{aligned}$$

Since the observed $T_{MLOD}$ score is greater than 3 this is taken as evidence of linkage between the marker and disease susceptibility loci.

### 2.1.2 Some Other Tests for Linkage

Another test statistic that was proposed by Haseman and Elston (1972) uses the total number of pairs of alleles IBD in affected sib pairs. In each sib pair that has two alleles IBD, there are two pairs of alleles that are IBD. There is one pair of alleles IBD in sib pairs that have one allele IBD, so the total number of allele pairs that are IBD is $n_1 + 2n_2$.

Each sib pair has 2 allele pairs (maternal and paternal), so the total number of allele pairs is $2n$. As shown earlier, the probability that an allele pair in sibs is IBD, is one half under the null hypothesis of no linkage. Because all allele pairs are independent, the total number IBD has a binomial distribution with index $2n$ and probability $1/2$, so the exact P-value can be calculated as

$$p = P(n_1 + 2n_2 \geq n_{1obs} + 2_{n2obs}).$$

In large samples a normal approximation can be used, with mean $n$ and variance $2n(1/2)(1/2) = n/2$ can be used. The standardized statistic

$$T_2 = \frac{n_1 + 2n_2 - n}{\sqrt{n/2}} \tag{2.3}$$

can be referred to the standard normal distribution, giving an approximate P-value

$$P = 1 - \Phi(T_{2obs})$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Another test for linkage that can be used for any type of relatives is a goodness of fit test. The probability of relatives having 0, 1 or 2 alleles IBD are given in Table 2.2 and these can be used to calculate expected counts. The hypotheses are

$H_O$ : $p_0 = \alpha_{R0}$, $p_1 = \alpha_{R1}$, $p_2 = \alpha_{R2}$
$H_A$ : $p_0$, $p_1$, $p_2$ not as specified in $H_O$.

The goodness of fit test statistic is

$$T = \sum_{cells} \frac{(observed - expected)^2}{expected}. \tag{2.4}$$

For relatives where an IBD status of 2 is impossible, there are only two cells, otherwise there are three. For $n$ affected sib pairs, the expected counts are $n/4$, $n/2$ and $n/4$ for 0, 1 and 2 alleles IBD respectively, and the test statistic as

$$T_3 = \frac{(n_0 - n/4)^2}{n/4} + \frac{(n_1 - n/2)^2}{n/2} + \frac{(n_2 - n/4)^2}{n/4}$$

$$= \frac{1}{n}[4(n_0)^2 + 2(n_1)^2 + 4(n_2)^2 - n^2]. \tag{2.5}$$

The test statistic is approximately distributed as a $\chi^2$ with 2 or 1 degrees of freedom depending on the possibility of IBD status 2.

The log likelihood-ratio test could be used as well. The test statistic for affected sib pairs (looking only at the number who have 0 alleles IBD) is

$$\begin{aligned} T_{LR} &= -2\ln\left[\frac{l(observed)}{l(null)}\right] \\ &= -2\ln\left[\frac{\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n-n_0}{n}\right)^{n-n_0}}{(1/4)^{n_0}(3/4)^{n-n_0}}\right] \\ &= -2\ln\left[\left(\frac{4n_0}{n}\right)^{n_0}\left(\frac{4(n-n_0)}{3n}\right)^{n-n_0}\right] \\ &= -2\left[n_0\ln\left(\frac{4n_0}{n}\right) + (n-n_0)\ln\left(\frac{4(n-n_0)}{3n}.\right)\right] \end{aligned} \tag{2.6}$$

The likelihood ratio statistic is approximately distributed as a $\chi^2$ with 2 or 1 degrees of freedom.

## 2.2 Power of Linkage Tests with ASPs

The power of linkage tests is complicated and depends on several genetic concepts that have not been discussed to this point. The most important of these genetic concepts is the recombination fraction, $\theta$, which is a measure of the linkage between two loci. The **recombination fraction** is the probability that a recombination occurs between the marker and disease loci. A recombination occurs during meiosis, when the chromosomes of an individual "cross-over" during replication. As mentioned earlier, each individual has two distinct copies of each chromosome and one of these chromosomes is passed on to the offspring of that individual. However, the creation of gametes is not a perfect process and sometimes a crossing over can occur. Ignoring the complicated biological process that results in this situation, a simple figure (Fig. 2.1, taken with permission from www.accessexcellence.org/AB/GG/crossing.html) will make the results of this clear.

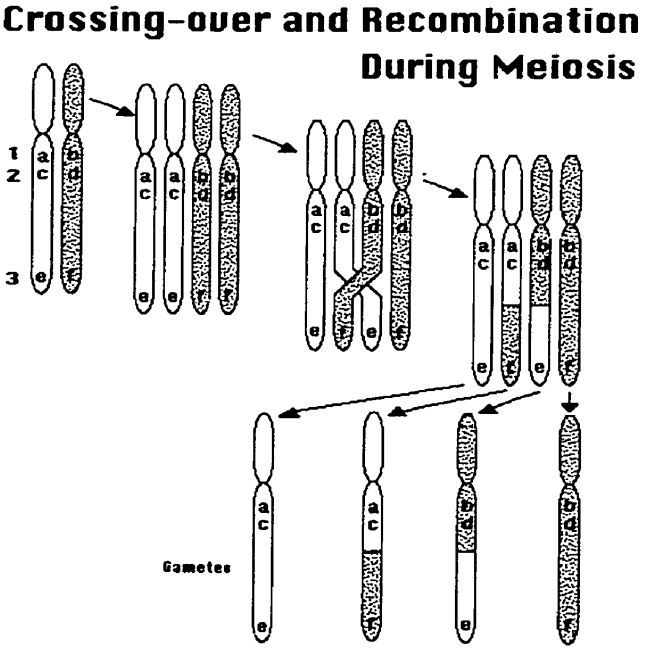


Figure 2.1: Crossing Over During Gamete Production

Both the light and dark chromosomes are replicated during the first stage of meiosis. Then the inner pair of chromosomes may cross over and recombine, forming two new chromosomes, each consisting of parts of the light and dark chromosomes. In the figure, recombination occurred between the loci indicated by the numbers 2 and 3. It should be noted that there was no recombination on one half of the gametes, thus $\theta$ is less than or equal to $1/2$. Loci on separate chromosomes are assumed to have a recombination fraction of $1/2$, since the transmission of these alleles is assumed to be independent. Recombination occurs less frequently between loci that are close together on the same chromosome, so $\theta$ can be used as a measure of distance between two loci. Two loci are said to be linked if their recombination fraction is less that $1/2$. Although there may be more than one crossing over on a single chromosome, this is an extremely rare event and will be ignored. It will also be assumed that the recombination fraction is equal for both sexes.

The power of linkage tests also depends on the nature of the disease. Diseases with a genetic component can depend on a single locus, or many loci and often do not follow simple dominance or recessive patterns.

Risch (1990a) studied affected relative pairs (allowing other relatives besides siblings) and expressed the power of these tests in terms of the sample size $n$, the recombination fraction $\theta$, and the relative risk ratio $\lambda_R$, which quantifies the increase in risk for a type R relative given that their family member is affected. He assumed initially that the disease is affected only by one locus, which can have any number of alleles. He later extended his results to multiple loci.

To calculate the power by this method, there are some quantities that we must know or be given. These are population prevalence of the disease, K, and the recurrence risk of some relative types, $K_R$. Some of the types of relatives Risch (1990a) refers to are monozygotic twins (M), siblings (S), and parent/offspring (O). For example, $\lambda_M$ is the relative risk ratio for monozygotic twins. Although he does use other relations, these are the only ones required to calculate the power of tests using affected sib pairs, which is the focus of this thesis.

Consider two relatives and define the random variable $X_1$ to be 1 if individual 1 has the disease, and 0 otherwise. Similarly, define $X_2$ for a relative of type R. The population prevalence of the disease, K, is then equal to the expected value of $X_1$. Also, the recurrence risk, $K_R = E(X_2|X_1 = 1)$, is the probability of a type R relative having the disease, given that their type R relative does. The probability that the first individual and their type R relative are both affected is $K \times K_R = E(X_1X_2) = K^2 + Cov(X_1, X_2)$. It follows that the recurrence risk is

$$K_R = K + \frac{1}{K}Cov(X_1, X_2)$$

which is a formula that was first derived by James (1971).

The **relative risk ratio**, $\lambda_R$, for a type R relative is then

$$\lambda_R = \frac{K_R}{K} = 1 + \frac{1}{K^2}Cov(X_1, X_2). \tag{2.7}$$

The covariance is positive if there is a genetic component to the disease, leading to a relative risk ratio greater than one.

James (1971) also indicated that the covariance (for a single locus disease) in (2.7) can be written as

$$\lambda_R = 1 + \frac{1}{K^2}(2\theta_{xy}V_A + \Delta_{xy}V_D), \tag{2.8}$$

where $V_A$ and $V_D$ are the additive genetic variance and dominance genetic variance of the penetrance, and $\theta_{xy}$ and $\Delta_{xy}$ are measures of relatedness for the individuals. The penetrance is the probability of getting the disease given a specific genotype.

Formal definitions of additive genetic variance and dominance genetic variance are not required for this thesis, as they are only play a minor part in the calculation of power for affected relatives. Let it suffice to say that the total genetic variance (amount of variance in a trait that can be explained by genes rather than environment) is divided into additive genetic variance and dominance genetic variance. Additive genetic variance is the amount of the genetic variance that can be accounted for by the regression of the trait onto the gene content (number (0, 1 or 2) of a specific allele

in the genotype at a locus) and dominance genetic variance is the residual genetic variance. In this case, the trait is the probability of getting the disease (Lynch and Walsh 1998).

The measures of relatedness are the coefficient of coancestry, $\theta_{xy}$, and the coefficient of fraternity, $\Delta_{xy}$. If one randomly selects one allele (at a particular locus) from individuals $x$ and $y$, the probability that these alleles are IBD is the **coefficient of coancestry**, $\theta_{xy}$. The **coefficient of fraternity**, $\Delta_{xy}$, is the probability that both alleles at a locus are IBD in individuals $x$ and $y$. These probabilities are, from Lynch and Walsh (1998), shown in Table 2.3.

| Relationship | $\theta_{xy}$ | $\Delta_{xy}$ |
|---|---|---|
| Parent-offspring | 1/4 | 0 |
| Grandparent-grandchild | 1/16 | 0 |
| Half sibs | 1/8 | 0 |
| Full sibs | 1/4 | 1/4 |
| First Cousins | 1/16 | 0 |
| Monozygotic (identical) twins | 1/2 | 1 |

Table 2.3: Measures of Relatedness for Different Relative Types

Note that closer relatives have larger measures of relatedness, which leads to larger relative risk ratios.

Calculating the power of the test requires evaluation of the probability of having a particular IBD status, given that both sibs are affected. These probabilities depend on the relative risk ratio for the disease as shown by Risch (1990b). The simpler situation in which there is no recombination is discussed first.

The probability that two affected relatives share $i$ alleles IBD is denoted $z_{Ri}$. Using Bayes Rule, this can be written as

$$\begin{aligned} z_{Ri} &= P(IBD = i | 2 \text{ relatives Affected}) \\ &= \frac{P(IBD = i) P(2 \text{ relatives Affected} | \text{IBD} = i)}{P(2 \text{ relatives affected})}. \end{aligned} \tag{2.9}$$

In particular, the probability of an affected relative pair having 0 alleles IBD is

$$\begin{aligned} z_{R0} &= P(IBD = 0|2 \text{ relatives affected}) \\ &= \frac{P(IBD = 0)P(2 \text{ relatives affected}|\text{IBD} = 0)}{P(2 \text{ relatives affected})}. \end{aligned}$$

Given that the relatives have no alleles IBD at the marker locus, then they have no alleles IBD at the disease locus, because of the assumption of no recombination. Since they have no alleles IBD at the disease locus, they can be treated as unrelated as far as the disease is concerned, and thus they both face only the general risk that all individuals face. This allows us to write

$$\begin{aligned} z_{R0} &= \frac{\alpha_{R0}(K \times K)}{K \times K_R} \\ &= \frac{\alpha_{R0}}{\lambda_R} \end{aligned} \tag{2.10}$$

where, as before, $\alpha_{R0}$ is the null probability of type R relatives having an IBD status of 0.

The probability that the affected relatives have one or two alleles IBD can be calculated in a similar way. In particular

$$\begin{aligned} z_{R1} &= P(IBD = 1|2 \text{ relatives affected}) \\ &= \frac{P(IBD = 1)P(2 \text{ relatives affected}|\text{IBD} = 1)}{P(2 \text{ relatives affected})}. \end{aligned}$$

If the relative pair have one allele IBD, they are as genetically similar at the disease locus as a parent-offspring pair. Thus the relative risk ratio for parent-offspring pairs, $\lambda_O$, can be used to calculate the probability of two relatives being affected given that they have one allele pair IBD

$$\begin{aligned} z_{R1} &= \frac{\alpha_{R1}(K \times K_O)}{K \times K_R} \\ &= \alpha_{R1}\frac{\lambda_O}{\lambda_R}. \end{aligned} \tag{2.11}$$

Also, for IBD status 2,

$$\begin{aligned} z_{R2} &= P(IBD = 2|2 \text{ relatives affected}) \\ &= \frac{P(IBD = 2)P(2 \text{ relatives affected}|\text{IBD} = 2)}{P(2 \text{ relatives affected})}. \end{aligned}$$

If they have an IBD status of two, then the pair are as genetically similar as monozygotic twins and the relative risk ratio for monozygotic twins, $\lambda_M$, can be used to simplify the expression

$$\begin{aligned} z_{R2} &= \frac{\alpha_{R2}(K \times K_M)}{K \times K_R} \\ &= \alpha_{R2}\frac{\lambda_M}{\lambda_R}. \end{aligned} \tag{2.12}$$

The power of the tests described above could be evaluated as a function of $\lambda_R$, $\lambda_M$ and $\lambda_O$ using the alternative probabilities $z_{Ri}$. However, these results are of limited value because of the unrealistic assumption that there is no recombination.

If recombination is possible, it is necessary to consider the potential for the IBD status at the marker locus to not equal the IBD status at the disease locus, because of a recombination event between the two loci. Recall that the location of the disease locus is unknown, as is the IBD status there. Using $IBD_m$ and $IBD_d$ to denote the IBD status at the marker and disease locus respectively, it is possible to expand the general equation (2.9) for the case of recombination as

$$z_{Ri} = \frac{P(2 \text{ affected} \cap \text{IBD}_\text{m} = i)}{P(2 \text{ affected})}$$

$$= \frac{1}{P(2 \text{ affected})} \sum_{j=0}^{2} P(2 \text{ affected} \cap IBD_m = i \cap IBD_d = j)$$
$$= \frac{1}{P(2 \text{ affected})} \sum_{j=0}^{2} P(2 \text{ affected} \cap \text{IBD}_\text{m} = i | IBD_d = j) P(IBD_d = j).$$

Note that the disease status is conditionally independent of the IBD status at the marker locus given the IBD status at the disease locus. This is because it is the disease locus that actually determines the disease. Thus,

$$z_{Ri} = \frac{1}{P(2 \text{ affected})} \sum_{j=0}^{2} P(2 \text{ affected} | IBD_d = i) P(IBD_m = i | IBD_d = j) P(IBD_d = j).$$

The order of conditioning in the last two terms can be reversed, giving

$$\begin{aligned} z_{Ri} &= \frac{1}{P(2 \text{ affected})} \sum_{j=0}^{2} P(2 \text{ affected} | IBD_d = i) P(IBD_d = j | IBD_m = i) P(IBD_m = i) \\ &= \frac{P(IBD_m = i)}{P(2 \text{ affected})} \sum_{j=0}^{2} P(2 \text{ affected} | IBD_d = i) P(IBD_d = j | IBD_m = i) \\ &= \frac{\alpha_{Ri}}{K \times K_R} \sum_{j=0}^{2} P(2 \text{ affected} | IBD_d = i) P(IBD_d = j | IBD_m = i). \end{aligned} \quad (2.13)$$

Haseman and Elston (1972) introduced the parameter $\Psi$ to assist in calculating the probabilities $P(IBD_d = j | IBD_m = i)$. This parameter is the probability that the IBD status for a pair of alleles (maternal or paternal) is the same at the marker locus and the disease locus. The parameter is given by

$$\Psi = \theta^2 + (1 - \theta)^2 \quad (2.14)$$

and is the probability of there being two recombinations, $\theta^2$, plus the probability of there being no recombination, $(1 - \theta)^2$. Note that it has been assumed that only one recombination is possible between the marker and disease locus, thus if there

are two recombinations, there is a single recombination event in the chromosome of each offspring. If there are no recombinations between the marker locus and the disease locus then the IBD status at the two loci are the same. Also, if there is a recombination on the same chromosome (maternal or paternal) in each individual then the IBD status is the same at the marker and disease locus. An example will help clarify this.

Consider the situation of ASPs, where a single parent is heterozygous at both the disease (alleles $D_1$ and $D_2$) and the marker locus (alleles $M_1$ and $M_2$), with $M_1$ and $D_1$ being on the same chromosome in the parent. Also, consider the offspring as being IBD at the marker chromosome for the allele from this parent (both have allele $M_1$). If there is no recombination (which happens with a probability of $(1-\theta)^2$) then both offspring have the $D_1$ allele at the disease locus and thus they are IBD at both the marker and disease locus. If there is a recombination in both offspring (which happens with a probability of $\theta^2$, then both offspring would have the $D_2$ allele at the disease locus, and thus would still have the same IBD status at the disease and marker loci.

Using $\Psi$, we can calculate the probability of affected relatives having $j$ alleles IBD at the disease locus given that have have $i$ alleles IBD at the marker locus. If sibs have 2 alleles IBD at the marker locus, the probability that they have 2 alleles IBD at the disease locus is $P(IBD_d = 2|IBD_m = 2) = \Psi^2$, as the two chromosomes (maternal and paternal) are considered independent of each other and the probability that they have the same IBD status at the disease locus is $\Psi$ for each chromosome. Similarly, the probability that sibs with two alleles IBD at the marker locus have no alleles IBD at the disease locus is $P(IBD_d = 0|IBD_m = 2) = (1-\Psi)^2$, since the probability that the IBD status changes on each independent chromosome is $1-\Psi$. Since the sum of all probabilities must add to one and the only other option is having one allele pair IBD at the disease locus, the probability of this is

$$P(IBD_d = 1|IBD_m = 2) = 1 - \Psi^2 - (1-\Psi)^2 = 2\Psi - 2\Psi^2 = 2\Psi(1-\Psi).$$

The probabilities for the other IBD statuses can be calculated in a similar fashion, with the arguments for the situation in which $IBD_m = 0$ the same as those for when $IBD_m = 2$. When the IBD status at the marker is one, for there to be two allele pairs IBD at the disease locus, there has to be a change in the IBD status on the chromosome in which the sibs are not IBD at the marker locus (which occurs with probability $\Psi$) and no change in the IBD status for the locus in which they are IBD (probability $1 - \Psi$). Since these events are independent, we can write $P(IBD_d = 2|IBD_m = 1) = \Psi(1 - \Psi)$.

The probabilities for sibs and some other relatives are summarized in Table 2.4 (Risch 1990b).

| Relationship | Marker IBD Status | Disease IBD Status 0 | 1 | 2 |
|---|---|---|---|---|
| Sibs | 2 | $\Psi^2$ | $2\Psi(1-\Psi)$ | $(1-\Psi)^2$ |
| | 1 | $\Psi(1-\Psi)$ | $\Psi^2+(1-\Psi)^2$ | $\Psi(1-\Psi)$ |
| | 0 | $(1-\Psi)^2$ | $2\Psi(1-\Psi)$ | $\Psi^2$ |
| Grandparent-grandchild | 1 | 0 | $1-\theta$ | $\theta$ |
| | 0 | 0 | $\theta$ | $1-\theta$ |
| Uncle-nephew | 1 | 0 | $\Psi(1-\theta)+\theta/2$ | $1-\theta/2-\Psi(1-\theta)$ |
| | 0 | 0 | $1-\theta/2-\Psi(1-\theta)$ | $\Psi(1-\theta)+\theta/2$ |
| Half-sibs | 1 | 0 | $\Psi$ | $1-\Psi$ |
| | 0 | 0 | $1-\Psi$ | $\Psi$ |
| First Cousins | 1 | 0 | $\Psi(1-\theta)^2+\frac{1}{2}\theta^2$ | $1-\frac{1}{2}\theta^2-\Psi(1-\theta)^2$ |
| | 0 | 0 | $1-\frac{1}{2}\theta^2-\Psi(1-\theta)^2$ | $\Psi(1-\theta)^2+\frac{1}{2}\theta^2$ |

Table 2.4: Conditional IBD Probabilities, given Marker IBD Status

The probability of an affected relative pair having $i$ allele pairs IBD at the marker locus is calculated using (2.13) and the appropriate entry in Table 2.4.

For affected sib pairs,

$$\begin{aligned} z_{S0} &= \frac{\alpha_{S0}}{K \times K_S}[K^2\Psi^2 + KK_O 2\Psi(1-\Psi) + KK_M(1-\Psi)^2] \\ &= \frac{1}{4\lambda_S}[\Psi^2 + \lambda_O 2\Psi(1-\Psi) + \lambda_M(1-\Psi)^2]. \end{aligned} \tag{2.15}$$

The relative risk ratio can be expressed in terms of the additive and dominance variance and measures of relatedness, as shown in (2.8). In particular $\lambda_M = 1 + (1/K)^2(V_A + V_D)$, $\lambda_S = 1 + (1/K)^2(V_A/2 + V_D/4)$ and $\lambda_O = 1 + (1/K)^2 V_A/2$. The risk ratio for monozygotic twins, $\lambda_M$, can be written in terms of $\lambda_S$ and $\lambda_O$

$$\lambda_M = 4\lambda_S - 2\lambda_O - 1. \tag{2.16}$$

Substituting (2.16) into (2.15) gives the $IBD_m = 0$ probability in terms of $\theta$, $\lambda_O$ and $\lambda_S$

$$\begin{aligned} z_{S0} &= \frac{1}{4\lambda_S}[\Psi^2 + \lambda_O 2\Psi(1-\Psi) + (4\lambda_S - 2\lambda_O - 1)(1-\Psi)^2] \\ &= \frac{1}{4\lambda_S}[4\lambda_S\Psi^2 - 4\lambda_O\Psi^2 - 8\lambda_S\Psi + 6\lambda_O\Psi + 4\lambda_S - 2\lambda_O + 2\Psi - 1] \\ &= \frac{1}{4\lambda_S}[\lambda_S + 2(\lambda_S - \lambda_O)(2\Psi^2 - 3\Psi + 1) - 2\Psi\lambda_S + 2\Psi + \lambda_S - 1] \\ &= \frac{1}{4} + \frac{1}{4\lambda_S}[2(\lambda_S - \lambda_O)(2\Psi - 1)(\Psi - 1) - (2\Psi - 1)(\lambda_S - 1)] \\ &= \frac{1}{4} - \frac{1}{4\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2(\Psi - 1)(\lambda_O - \lambda_S)]. \end{aligned} \tag{2.17}$$

Similarly for $IBD_m = 1$

$$\begin{aligned} z_{S1} &= \frac{1/2}{K \times K_S}[K^2\Psi(1-\Psi) + KK_O[\Psi^2 + (1-\Psi)^2] + KK_M\Psi(1-\Psi)] \\ &= \frac{1}{2\lambda_S}[\Psi(1-\Psi) + \lambda_O[\Psi^2 + (1-\Psi)^2] + \lambda_M\Psi(1-\Psi)] \\ &= \frac{1}{2\lambda_S}[\lambda_S(-4\Psi^2 + 4\Psi) + \lambda_O(4\Psi^2 - 4\Psi + 1)] \\ &= \frac{1}{2} - \frac{1}{2}(2\Psi - 1)^2\frac{1}{\lambda_S}(\lambda_S - \lambda_O), \end{aligned} \tag{2.18}$$

and for $IBD_m = 2$

$$\begin{aligned} z_{S2} &= \frac{1/4}{K \times K_S}[K^2(1-\Psi)^2 + KK_O 2\Psi(1-\Psi) + KK_M\Psi^2] \\ &= \frac{1}{4\lambda_S}[(1-\Psi)^2 + \lambda_O 2\Psi(1-\Psi) + \lambda_M\Psi^2] \\ &= \frac{1}{4\lambda_S}[-2\Psi + 1 - \lambda_O(4\Psi^2 - 2\Psi) + \lambda_S(4\Psi^2 - 2\Psi) + (2\Psi\lambda_S - \lambda_S) + \lambda_S] \\ &= \frac{1}{4} + \frac{1}{4\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2\Psi(\lambda_S - \lambda_O)]. \end{aligned} \tag{2.19}$$

Note that these probabilities equal the null values of 1/4, 1/2 and 1/4 if $\theta = 1/2$ (so $\Psi = 1/2$), or if there is no elevated risk of the disease for sibs or parent/offspring pairs, $\lambda_S = \lambda_O = 1$.

Risch (1990b) then uses two arguments to simplify these formulas and to better justify the test statistic based on $n_0$ given in (2.2). The first argument is that if $\theta$ is near zero then $\Psi$ is close to one and $\Psi - 1$ is close to zero. Thus, for $\theta$ near zero,

$$z_{S0} = \frac{1}{4} - \frac{1}{4}(2\Psi - 1)\frac{1}{\lambda_S}(\lambda_S - 1) \tag{2.20}$$

which depends only on the recombination fraction $\theta$ (through $\Psi$) and the risk ratio for siblings, $\lambda_S$. Both $z_{S1}$ and $z_{S2}$ still depend on $\theta$, $\lambda_S$ and the risk ratio for parents/offspring, $\lambda_O$, when $\theta$ is close to zero. Thus, when $\theta$ is close to zero, the power for tests that use only $n_0$, the number of sib pairs that have no allele pairs IBD, is dependent on only one parameter, $\lambda_S$, while the power for tests that use the number of sib pairs that have one or two allele pairs IBD is dependent on two parameters, $\lambda_S$ and $\lambda_O$. This is a considerable simplification.

Risch (1990b) does concede that for diseases that have a significant dominance variance component (such as a rare recessive disorder), a test that uses $n_2$ will be more powerful, as $z_{S2}$ can approach 1, whereas $z_{S0}$ approaches zero. The null probabilities are 1/4 for both probabilities and so there is a larger departure from the null using $n_2$. However, he mentions that in the absence of a dominance variance component $\lambda_S = \lambda_O$, and so $z_{S1} = 1/2 = z_{S0} + z_{S2}$. In this case,

$$z_{S0} = \frac{1}{4} - \frac{1}{4}(2\Psi - 1)\frac{1}{\lambda_S}(\lambda_S - 1) \tag{2.21}$$

and

$$z_{S2} = \frac{1}{4} + \frac{1}{4}(2\Psi - 1)\frac{1}{\lambda_S}(\lambda_S - 1) \tag{2.22}$$

so tests based on the observed value of $z_{S0}$ or $z_{S2}$ are equivalent.

Risch (1990b) then says "most common complex diseases in man show little or no dominance effect, i.e. $\lambda_S$ and $\lambda_O$ are similar. This is true, for example, for common cancers, cardiovascular disease, psychiatric disorders, birth defects and so on." (Risch 1990b) He uses this as a basis to justify both the fact that he only examines in detail the case in which $\lambda_S = \lambda_O$, and the use of the test statistic based on $n_0$.

Another benefit of using $n_0$ instead of $n_2$ is the fact that the test can be generalized to other relative types. Risch (1990b) examined relative types other than sibs, and since siblings are the only relation that can possibly have two alleles identical by descent, a test statistic based on the number of sibs that have two alleles IBD could not be generalized.

There is considered to be evidence of linkage when the $T_{MLOD}$ score is greater than or equal to 3. This is equivalent to $n_0 \leq W$ for some W which depends on $n$. Referring back to (2.2), this W satisfies

$$\begin{aligned} 3 &\leq W \log_{10}\left(\frac{4W}{n}\right) + (n-W)\log_{10}\left(\frac{4(n-W)}{3n}\right) \\ 10^3 &\leq \left(\frac{4W}{n}\right)^W \left(\frac{4(n-W)}{3n}\right)^{(n-W)} . \end{aligned} \tag{2.23}$$

A rather simple method was used to calculate the value W. Starting with the whole number $k$ less than $n/4$, and for each integer less than $k$ the likelihood ratio statistic given by the right hand side of (2.23) was calculated. The largest integer giving a value greater than 1000 ($10^3$) is W for that value of $n$. For $n = 50$, $W = 2$; for $n = 100$, $W = 10$; for $n = 200$, $W = 28$; and for $n = 300$, $W = 40$.

The power is calculated as $P(n_0 \leq W)$ using the binomial distribution with index $n$ and probability $z_{S0}$. This probability is calculated from (2.20) using choices for the recombination fraction $\theta$ and risk ratio $\lambda_S$. Power curves are shown in Fig. 2.2 for the case in which there is no recombination and in Fig. 2.3 for different values of $\theta$.

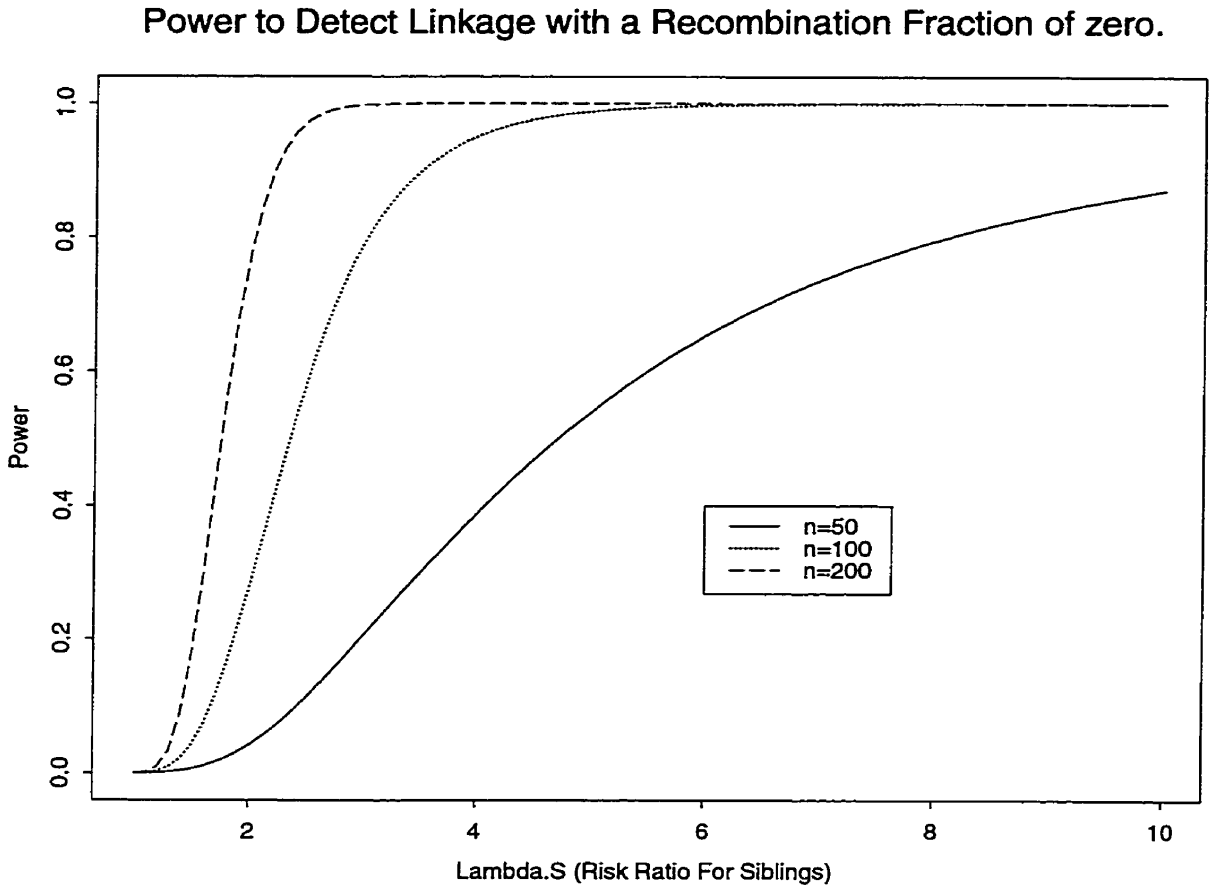


Figure 2.2: Power to Detect Linkage with the MLOD Test for different values of $n$

As would be expected, it is easier to detect linkage for diseases with a large relative risk ratio and with very tight linkage ($\theta$ close to zero). A plot of power for several values of $n$, Figure 2.2, shows more clearly the association between sample size, relative risk ratio and power.

A plot of how the recombination fraction affects the power is also useful (Figure 2.3). As can be seen the power of the test decreases rapidly for larger values of $\theta$. This makes sense because recombination implies weaker linkage between the marker and disease loci.

It should be noted that the sample sizes used can be thought of a minimum sample sizes that would be needed to obtain a certain level of power. This is because

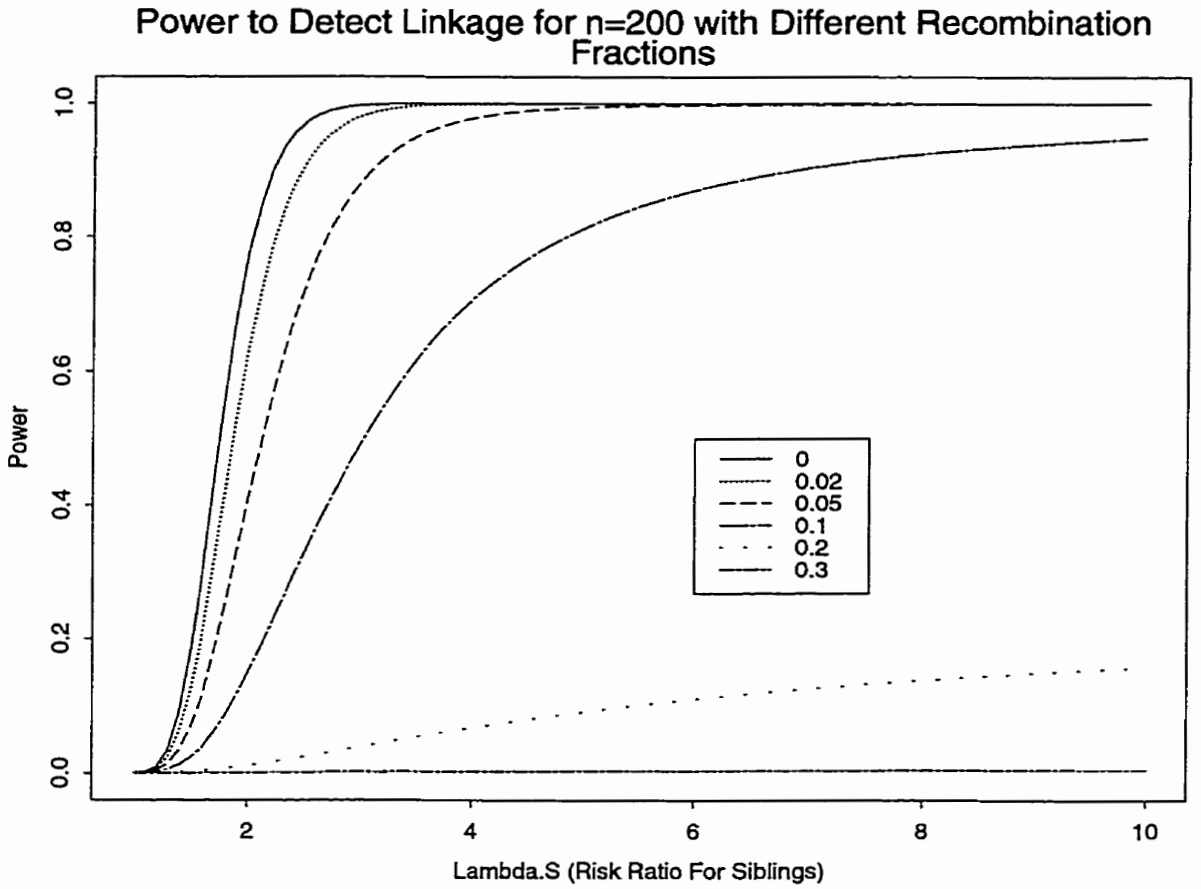


Figure 2.3: Power to Detect Linkage with the MLOD Test for different values of $\theta$

we have assumed that the disease was influenced only by one locus and there was no significant dominance effect. If the disease was influence by more than one locus, we could expect that each locus would contribute less to the relative risk ratio and thus be harder to detect.

It is also possible to use the method developed by Risch (1990b) to calculate the power of the test that uses only $n_2$, given in (2.1). Risch (1990b) gave a formula for $z_{S2}$ when the dominance variance is assumed to be insignificant (2.22). The power is $P(n_2 \geq c)$ for a critical value $c$ which depends on the sample size $n$ and the significance level $\alpha$. This probability can be evaluated exactly using a binomial distribution with index $n$ and probability $z_{S2}$.

The $T_1$ test (2.1), which is the normal approximation, will be significant when $T_1$ is greater than $z_\alpha$ (the critical value for a one sided normal test). However in the MLOD test an approximate significance level of 0.0001 is used which corresponds to a one sided normal score of 3.72. To reasonably compare the power of the two tests, it is necessary to use the same level of significance for both tests. For a given $n$, the critical value satisfies

$$\begin{aligned} \frac{c - n/4}{\sqrt{3n/16}} &= 3.72 \\ c &= 3.72\sqrt{\frac{3n}{16}} + \frac{n}{4} \\ c &= 1.61\sqrt{n} + \frac{n}{4}. \end{aligned}$$

The power, $\mathrm{P}(n_2 \geq c)$, is then calculated using the binomial distribution with index $n$ and probability $z_{S2}$, for $z_{S2}$ given by (2.22) for some choice of $\theta$ and $\lambda_S$.

Power curves are shown in Figure 2.4 for the situation in which there is no recombination and in Figure 2.5 for different values of $\theta$ with $n = 200$. As can be seen from the plots, this test offers less power than the MLOD test. This is surprising since Risch (1990b) stated that tests based on the observed value of $z_{S0}$ and $z_{S2}$ are equivalent. To test this statement, it is possible to modify the MLOD test to examine $n_2$ instead of $n_0$ for affected sibs. The test statistic is

$$T_{MLOD2} = n_2 \log_{10}\left(\frac{n_2}{n\alpha_{S2}}\right) + (n - n_2)\log_{10}\left(\frac{n - n_2}{n - n\alpha_{S2}}\right) \tag{2.24}$$

and the hypotheses are

$H_O : p = \alpha_{S2}$
$H_A : p > \alpha_{S2}.$

The power can be calculated as before and a power curve is shown in Figure 2.6.

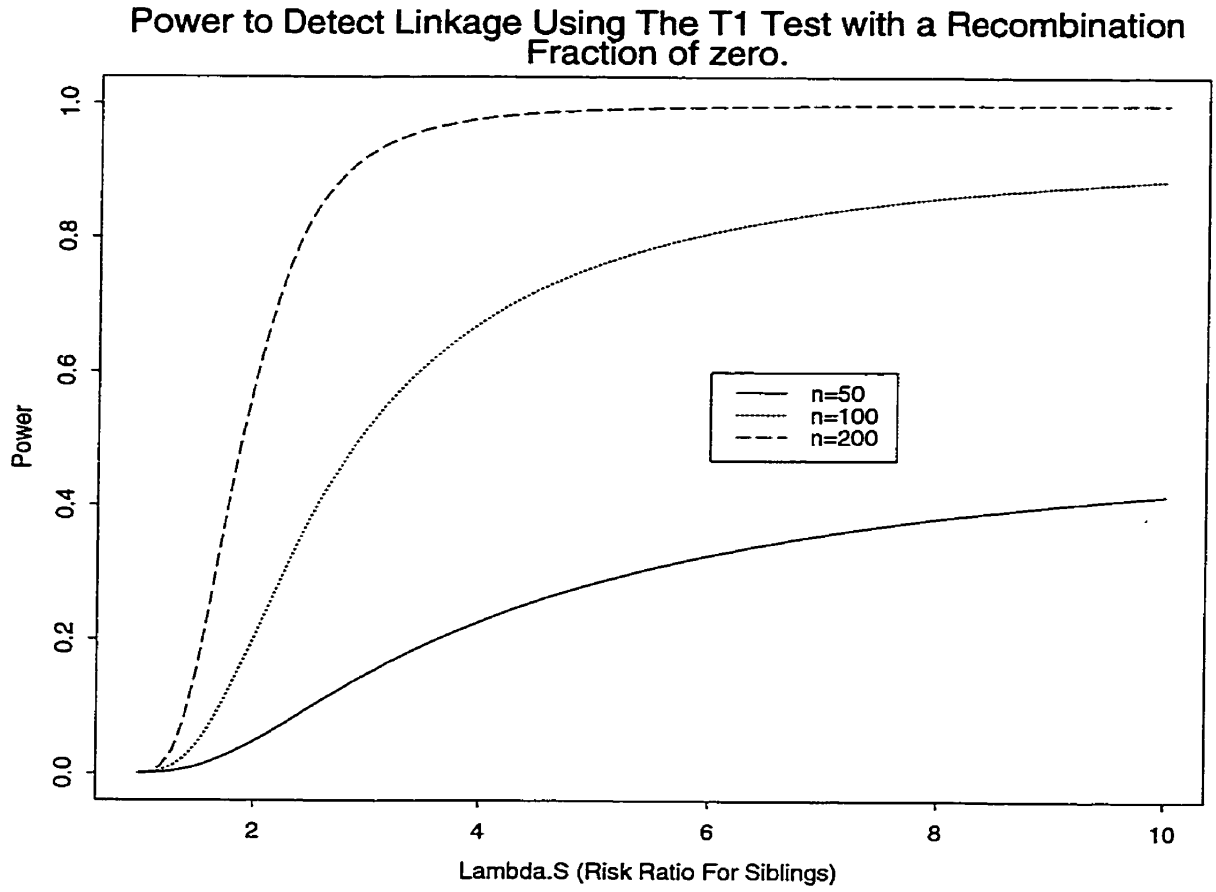


Figure 2.4: Power to Detect Linkage with the $T_1$ Test for different values of $n$

Careful inspection of Figure 2.6 and Figure 2.3 show that the modified test actually has less power despite using identical testing criteria, which contradicts Risch's statement about the equivalence of tests based on $n_0$ and $n_2$. This is a result of the fact that the variance of a binomial distribution is dependent on the probability $p$ and the variance decreases as $p$ moves away from 1/2. Although the formulas for the probabilities $z_{S0}$ and $z_{S2}$ given in (2.21) and (2.22) are similar, they will give probabilities that result in very different variances. Since $z_{S1} = 1/2$ under the assumptions of no significant dominance variance, $z_{S0} + z_{S2} = 1/2$ and $z_{S0}$ will always be closer to 0 if there is a genetic effect and thus a binomial distribution with a probability $z_{S0}$ have less variance than one with probability $z_{S2}$. This is what results in the difference

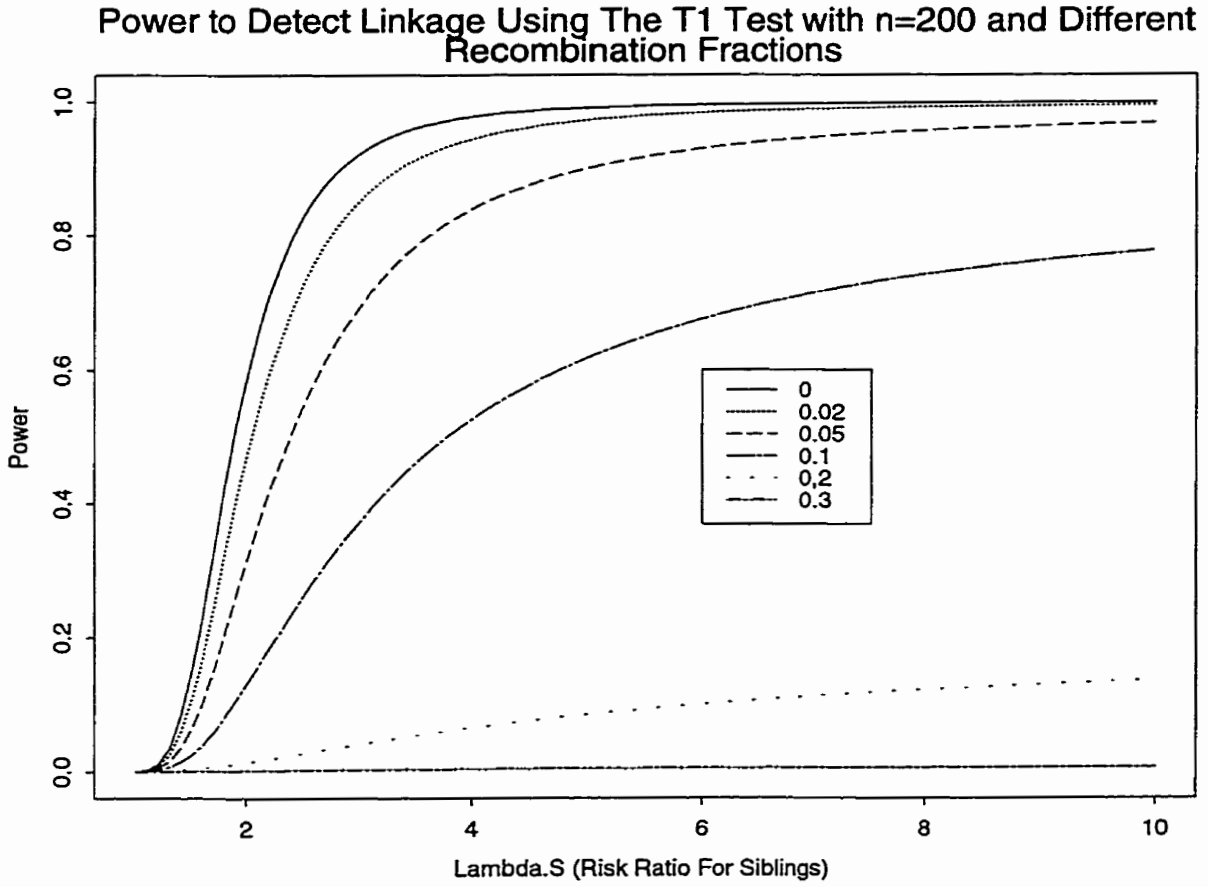


Figure 2.5: Power to Detect Linkage with the $T_1$ Test for different values of $\theta$

in power between these tests, as under the alternative hypothesis the values $n_0$ and $n_2$ come from binomial distributions with probabilities $z_{S0}$ and $z_{S2}$ respectively and deviations will be easier to detect when there is a smaller variance.

However, if there is a significant dominance variance component, we would expect tests based on $n_2$ to be more powerful since, as mentioned earlier, $z_{S2}$ can approach 1 if there is a significant dominance variance component.

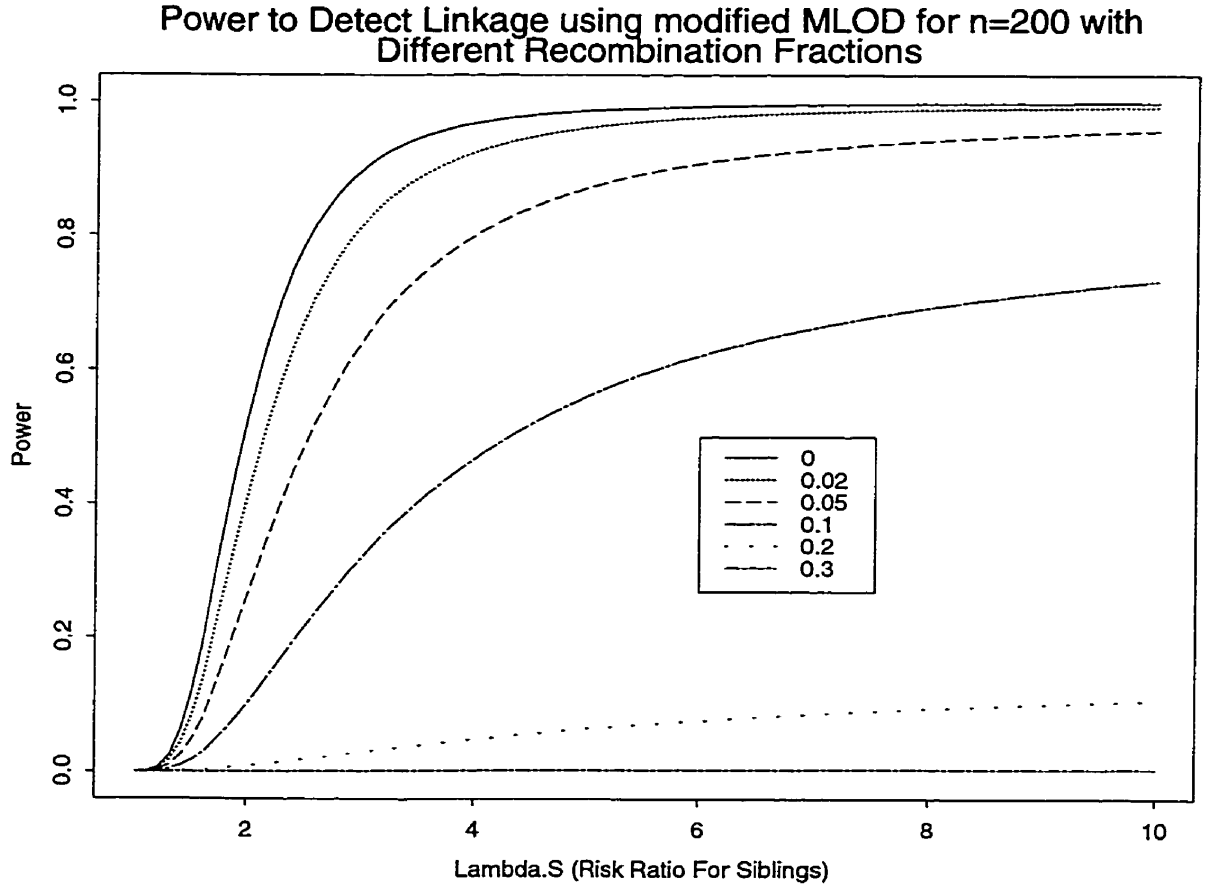


Figure 2.6: Power to Detect Linkage with the Modified MLOD Test for $n = 200$ with different values of $\theta$

# Chapter 3

# Linkage Disequilibrium and the Transmission Disequilibrium Test

As shown earlier, association between an allele at a marker locus and a disease does not imply that the allele associated with the disease is related to the disease, nor does it necessarily correspond to an increased probability of the disease. Finding a marker locus that is linked to a disease gives evidence that the marker locus is close to the disease locus on the genome. However, linkage tests are only concerned with the transmission of alleles to affected individuals and not the alleles that affected individuals have. They tell us nothing about the probability of having the disease for a particular marker genotype.

Ideally, it would be best to know that a marker locus is closely linked to the disease and that there is association with a particular allele. The perfect situation is a recombination fraction of 0 and strong association, in which case the marker locus is either extremely close to the disease locus or is the disease locus itself. Either way, having the associated allele tells you something about the probability of having the disease.

If there is both linkage and association between a marker locus and a disease locus there is said to be **linkage disequilibrium** or **gametic phase disequilibrium**. This arises when the transmission of the disease alleles and marker alleles is not

independent (i.e. there is linkage) and one allele occurs more often in individuals who have the disease (association). Consider the situation in which there is a disease locus with alleles $D_1$ and $D_2$ at the disease locus and a marker locus with alleles $M_1$ and $M_2$. Also assign population probabilities of the $D_1$ and $M_1$ alleles as $p$ and $m$ respectively. If the transmission of these alleles is independent, then the probability that an individual has both alleles $M_1$ and $D_1$ on the same chromosome is $mp$. However, if they are not independent (there is an association between the alleles at the disease and marker locus), the $D_1$ and $M_1$ alleles will occur together on the same chromosome with a different frequency.

We can define the **coefficient of disequilibrium**, $\delta$, to be $\delta = P(M_1D_1) - mp$ and thus $\delta$ is zero if the alleles at the marker and disease locus are independent. It is important to note that if there is no association, but there is linkage, the alleles on a chromosome are still independent but the transmission of alleles from a specific individual is not independent.

It should also be noted that in the previously discussion of association, the focus was on disease-allele association in which a particular allele is associated with the disease. Now the focus if on whether there is allelic association between the marker and disease *locus*, which means that certain alleles occur together in individuals more or less often than expected. Generally, if there is disease-allele association there will be allelic association between the alleles on the marker and disease loci.

It is possible to test for linkage at a marker locus that is known to have a disease-allele association and use the fact that there is association to increase the power of the test for linkage. Like most linkage tests, the transmission of alleles will be examined, but unlike most tests for linkage, the transmission of alleles from parents to a single affected offspring are considered. This eliminates the problem of population stratification in association tests and it is easier to obtain subjects than other tests for linkage.

Consider the situation in which there is an allele, $M_1$, that is associated with the disease of interest and the genotypes of the $2n$ parents of $n$ affected individuals are

known. Examining the alleles that the parents transmit and do not transmit to the affected offspring at the marker locus, the data could be represented as Table 3.1.

| Transmitted Allele | Nontransmitted Allele $M_1$ | Nontransmitted Allele $M_2$ | Total |
|---|---|---|---|
| $M_1$ | $a$ | $b$ | $a+b$ |
| $M_2$ | $c$ | $d$ | $c+d$ |
| Total | $a+c$ | $b+d$ | $2n$ |

Table 3.1: Counts of Transmitted and Nontransmitted Marker Alleles Among $2n$ Parents of $n$ Affected Children

If there is linkage and association between the marker and disease loci, the positively associated allele should be transmitted to affected offspring more often than would be expected if there was no linkage. This preferential transmission of the positively associated allele forms the basis of the **transmission/disequilibrium test** or **TDT** developed by Spielman et al. (1993). If the parents are homozygous, they do not provide any information about preferential transmission of alleles, since they only have one allele to transmit. Only parents heterozygous for the associated marker allele provide information about preferential transmission and only those parents will be used. Therefore, the only values of interest in Table 3.1 are $b$ and $c$.

Let $T(X)$ indicate that marker allele X is transmitted to the offspring and let $\bar{T}(Y)$ indicate that the marker allele Y is not transmitted to the offspring. Furthermore, let A be the event that the parent has an affected offspring and H be the event that the parent is heterozygous. The hypotheses of interest are

$H_O : \pi = 1/2$
$H_A : \pi \neq 1/2$

where $\pi = P(T(M_1) \cap \bar{T}(M_2)|A \cap H)$ is the probability that a heterozygous parent transmits $M_1$ and not $M_2$ to an affected child. This is a McNemar test, and the

P-value can be calculated exactly using the binomial distribution or approximatly using a $\chi^2$ or a normal distribution. The number $b$ in Table 3.1 follows a binomial distribution with index $b+c$ and probability $\pi = 1/2$ under the null hypothesis. Therefore the quantity

$$\begin{aligned} \chi^2_{TDT} &= \left(\frac{b-(b+c)/2}{\sqrt{(b+c)/4}}\right)^2 \\ &= \frac{(b-c)^2}{(b+c)} \end{aligned} \tag{3.1}$$

has an approximate $\chi^2$ distribution with 1 degree of freedom.

The test based on (3.1) is known as the "transmission/disequilibrium test" or TDT and was developed by Spielman et al. (1993). It is used as a test for linkage between the marker and disease loci, but as will be shown, there must be association in order to detect linkage. Spielman et al. (1993) also showed that the test can be used to detect both linkage and association. The TDT is also not hampered by some of the usual restrictions on linkage tests, such as the need for affected sib pairs with informative parents. The TDT uses information on families that have at least one affected offspring and at least one parent who is heterozygous for the allele that shows association. If both parents are heterozygous for the associated marker allele they can both be used because their transmission of alleles is independent under the null hypothesis. When you consider the fact that for most tests for linkage you need two affected siblings and parents that allow you to uniquely determine their IBD status, it is clear that it is easier to obtain subjects for the TDT test. This is true for both early-onset diseases, since parents who have a child with a disease that has a genetic influence may be hesitant to have another child, and late-onset disease, as only one parent has to be typed for the disease.

If a family has more than one affected offspring, each offspring can be used with the TDT. Once again, this is because the transmission of alleles is independent under

the null hypothesis of no linkage. The TDT with affected sibs will be further examined in Section 3.3.

The probability $\pi$ in the hypotheses above depends on the marker and disease allele probabilities, $m$ and $p$, and on the recombination fraction, $\theta$, and the disequilibrium coefficient, $\delta$, as well as the mode of inheritance for the disease. It is possible to find expressions for $\pi$ for different modes of inheritance, but for simplicity it is assumed that the disease is recessive and caused by the $D_1$ allele. Letting G denote the genotype of the parent on both chromosomes at the marker and disease loci, it is possible to use the law of total probability to write

$$\gamma_{ij} = P(T(M_i) \cap \bar{T}(M_j)|A) = \sum_G P(T(M_i) \cap \bar{T}(M_j)|G \cap A)P(G|A) \tag{3.2}$$

where $\gamma_{ij}$ if the probability that a parent transmits the $M_i$ allele and does not transmit the $M_j$ allele to an affected offspring, unconditional on the parent being heterozygous. The conditional genotype probabilities, $P(G|A)$, can be obtained using Bayes formula

$$P(G|A) = \frac{P(A|G)P(G)}{P(A)}. \tag{3.3}$$

Under random mating, the unconditional genotype probabilities are the product of the **haplotype** (combinations of alleles at the marker and disease loci on a chromosome) probabilities. The possible haplotypes are $M_1D_1$, $M_1D_2$, $M_2D_1$ and $M_2D_2$. If there is linkage disequilibrium, the allele at the disease locus is not independent of the allele at the marker locus. The probabilities of the four haplotypes are $\mathrm{P}(M_1D_1) = mp+\delta = x_1$, $\mathrm{P}(M_1D_2) = m(1-p)-\delta = x_2$, $\mathrm{P}(M_2D_1) = (1-m)p-\delta = x_3$ and $\mathrm{P}(M_2D_2) = (1-m)(1-p)+\delta = x_4$. (Using this notation, the coefficient of disequilibrium is $\delta = x_1x_4 - x_2x_3$).

The probability of an affected child (ignoring the genotype of the parent), $P(A)$, is simply $p^2$, the probability that the child gets two disease alleles, because of the assumption of a recessive disease.

The last thing needed to calculate the conditional genotype probability (3.3) is $P(A|G)$, the probability that a child is affected given the particular genotype for the

parent. There are only three possibilities for this probability. If the genotype of the parent does not involve any disease alleles (i.e. the parent does not have any $D_1$ alleles), then the child could not get the disease, since the child has to get a disease allele from both parents in order to get the disease. Thus if the particular genotype has no $D_1$ alleles then P($A$|No $D_1$ Alleles) = 0. If the parent had two $D_1$ alleles then the child will have the disease if it gets a disease allele from the other parent, which happens with a probability of $p$. Therefore, P($A|2\,D_1$ Alleles) = p. Also, if the parent has one $D_1$ allele, that allele will get passed with a probability of 1/2 and other parent will independently pass a disease allele with probability $p$ so P(A|1 $D_1$ Allele) = p/2.

As an example, consider the genotype $M_1D_1M_2D_2$. The conditional probability of this genotype given that they have an affected child is

$$\begin{aligned} P(M_1D_1M_2D_2|A) &= \frac{P(A|M_1D_1M_2D_2)P(M_1D_1M_2D_2)}{P(A)} \\ &= \frac{P(A|1\ D_1\ \text{Allele})\text{P}(\text{M}_1\text{D}_1)\text{P}(\text{M}_2\text{D}_2)}{P(A)} \\ &= \frac{(p/2)x_1x_4}{p^2} \\ &= \frac{x_1x_4}{2p}. \end{aligned}$$

Repeating this calculation for each of the sixteen (4 × 4) possible genotypes or combinations of haplotypes a parent could have gives Table 3.

The final probabilities to calculate in expression (3.2) is the transmission probability $P(T(M_1) \cap \bar{T}(M_2)|G \cap A)$ for each genotype. It is important to remember that the offspring is affected and so has two $D_1$ alleles, one passed from each parent.

Consider the simplest example for demonstration, in which the parent transmits $M_1$ and does not transmit $M_1$. The only way in which this occurs is if the parent has two $M_1$ marker alleles and thus we can restrict our attention to the top left quadrant

| | $M_1D_1$ | $M_1D_2$ | $M_2D_1$ | $M_2D_2$ | Sum |
|---|---|---|---|---|---|
| $M_1D_1$ | $x_1^2/p$ | $x_1x_2/2p$ | $x_1x_3/p$ | $x_1x_4/2p$ | $x_1(1+p)/2p$ |
| $M_1D_2$ | $x_1x_2/2p$ | 0 | $x_2x_3/2p$ | 0 | $x_2/2$ |
| $M_2D_1$ | $x_1x_3/p$ | $x_2x_3/2p$ | $x_3^2/p$ | $x_3x_4/2p$ | $x_3(1+p)/2p$ |
| $M_2D_2$ | $x_1x_4/2p$ | 0 | $x_3x_4/2p$ | 0 | $x_4/2$ |
| Sum | $x_1(1+p)/2p$ | $x_2/2$ | $x_3(1+p)/2p$ | $x_4/2$ | 1 |

Table 3.2: Conditional Genotype Distribution of One Parent Given that Parent has an Affected Child

of Table 3. Since both alleles are $M_1$, the probability that $M_1$ gets transmitted and $M_1$ is not transmitted is one for all genotypes with a disease allele in that quadrant.

All of these transmission probabilities will be 0, 1/2, 1, $\theta$, or $1-\theta$, depending on the marker alleles that are present and whether a recombination has to occur or not in order for the required marker allele to be transmitted with a disease allele. As an example consider $P(T(M_1) \cap \bar{T}(M_2)|M_1D_2M_2D_1 \cap A)$. Since the child is affected, we know that the $D_1$ allele was transmitted. $M_1$ would be transmitted with the disease allele, $D_1$, only if there was a recombination, which happens with probability $\theta$. Thus $P(T(M_1) \cap \bar{T}(M_2)|M_1D_2M_2D_1 \cap A) = \theta$.

It is now possible to calculate the transmission probabilities of equation (3.2) using Table 3 and the ideas in the last two paragraphs. For example,

$$\begin{aligned}
\gamma_{11} &= P(T(M_1) \cap \bar{T}(M_1)|A) \\
&= \sum_G P(T(M_1) \cap \bar{T}(M_1)|\mathrm{G} \cap \mathrm{A})\mathrm{P}(\mathrm{G}|\mathrm{A}) \\
&= 1P(M_1D_1M_1D_1|A) + 1P(M_1D_1M_1D_2|A) + 1P(M_1D_2M_1D_1|A) \\
&= \frac{x_1^2}{p} + \frac{x_1x_2}{2p} + \frac{x_1x_2}{2p} \\
&= \frac{x_1(x_1+x_2)}{p} \\
&= \frac{(mp+\delta)[mp+\delta+m(1-p)-\delta)]}{p} \\
&= m(m+\frac{\delta}{p})
\end{aligned}$$

which is the probability that a parent of an affected child transmits an $M_1$ allele and also does not transmit an $M_1$ allele.

It is more complicated to calculate $\gamma_{12}$. The symmetry of Table 3 can be used to simplify calculations, since the genotypes $M_iD_jM_kD_l$ and $M_kD_lM_iD_j$ are identical with respect to the alleles present on each chromosome. Thus

$$P(T(M_i) \cap \bar{T}(M_k)|M_iD_jM_kD_l \cap A) = P(T(M_i) \cap \bar{T}(M_k)|M_kD_lM_iD_j \cap A)$$

and

$$P(M_iD_jM_kD_l)|A) = P(M_kD_lM_iD_j|A)$$

as can be seen in Table 3. This allows us to write

$$\begin{aligned}
\gamma_{12} &= P(T(M_1) \cap \bar{T}(M_2)|A) \\
&= \sum_G P(T(M_1)\bar{T}(M_2)|\mathrm{G} \cap \mathrm{A})\mathrm{P}(\mathrm{G}|\mathrm{A}) \\
&= 2P(T(M_1) \cap \bar{T}(M_2)|M_1D_1M_2D_1 \cap A)P(M_1D_1M_2D_1|A) + \\
&\quad 2P(T(M_1) \cap \bar{T}(M_2)|M_1D_1M_2D_2 \cap A)P(M_1D_1M_2D_2|A) + \\
&\quad 2P(T(M_1) \cap \bar{T}(M_2)|M_1D_2M_2D_1 \cap A)P(M_1D_2M_2D_1|A) \\
&= 2\left(\frac{1}{2}\right)\frac{x_1x_3}{p} + 2(1-\theta)\frac{x_1x_4}{2p} + 2\theta\frac{x_2x_3}{2p} \\
&= \frac{x_1x_3 + x_1x_4 - \theta(x_1x_4 - x_2x_3)}{p} \\
&= \frac{x_1x_3 + x_1x_4 - \theta\delta}{p}.
\end{aligned}$$

Substituting the values of $x_1$, $x_2$, $x_3$ and $x_4$ expressed in terms of $m$, $p$ and $\delta$ gives

$$\gamma_{12} = \frac{(mp+\delta)[(1-m)p - \delta] + (mp+\delta)[(1-m)(1-p) + \delta] - \theta\delta}{p}$$

$$= \frac{1}{p}[(mp+\delta)(1-m)-\theta\delta]$$
$$= m(1-m)+\frac{1}{p}(1-m-\theta)\delta.$$

The probabilities $\gamma_{21}$ and $\gamma_{22}$ can also be calculated by this method and the probabilities are given in Table 3.3.

| Transmitted Allele | Nontransmitted Allele $M_1$ | Nontransmitted Allele $M_2$ | Total |
|---|---|---|---|
| $M_1$ | $m^2+m\frac{\delta}{p}$ | $m(1-m)+(1-\theta-m)\frac{\delta}{p}$ | $m+(1-\theta)\frac{\delta}{p}$ |
| $M_2$ | $m(1-m)+(\theta-m)\frac{\delta}{p}$ | $(1-m)^2-(1-m)\frac{\delta}{p}$ | $1-m-(1-\theta)\frac{\delta}{p}$ |
| Total | $m+(\theta\delta/p)$ | $1-m-(\theta\delta/p)$ | 1 |

Table 3.3: Probabilities of Combinations of Transmitted and Nontransmitted Marker Alleles For Parents of Affected Children

Inspection of Table 3.3 shows that the only values in Table 3.1 that have probabilities which depend on $\theta$ are $b$ and $c$, which supports the statement that only heterozygous parents provide information about linkage. Since only heterozygous parents are used, the responses will all fall in these two categories. Thus, the probability $\pi$ we are interested in is

$$\pi = P(T(M_1)\cap\bar{T}(M_2)|A\cap H)$$
$$= \frac{P(T(M_1)\cap\bar{T}(M_2)\cap H|A)}{P(H|A)}$$

The event $T(M_1)\cap\bar{T}(M_2)$ is a subset of the set of heterozygous parents, thus $P(T(M_1)\cap\bar{T}(M_2)\cap H|A) = P(T(M_1)\cap\bar{T}(M_2)|A)$. Also, the probability of a parent being heterozygous is the sum of the the probabilities of the two possible transmissions. Thus

$$\begin{aligned}\pi &= \frac{P(T(M_1) \cap \bar{T}(M_2)|A)}{P(T(M_1) \cap \bar{T}(M_2)|A) + P(T(M_2) \cap \bar{T}(M_1)|A)} \\ &= \frac{\gamma_{12}}{\gamma_{12} + \gamma_{21}}\end{aligned}$$

Taking the probabilities $\gamma_{12}$ and $\gamma_{21}$ from Table 3.3,

$$\begin{aligned}\pi &= \frac{m(1-m) + [(1-\theta-m)\delta/p]}{m(1-m) + [(1-\theta-m)\delta/p] + m(1-m) + [(\theta-m)\delta/p]} \\ &= \frac{m(1-m) + [(1-\theta-m)\delta/p]}{2m(1-m) + (1-2m)\delta/p}. \qquad (3.4)\end{aligned}$$

Substituting in $\theta = 1/2$ for the case of no linkage, gives

$$\begin{aligned}\pi &= \frac{m(1-m) + [(1/2-m)\delta/p]}{2m(1-m) + (1-2m)\delta/p} \\ &= \frac{m(1-m) + 1/2(1-m)\delta/p]}{2m(1-m) + (1-2m)\delta/p} = \frac{1}{2},\end{aligned}$$

which agrees with the null hypothesis as stated earlier. Also, when $\delta = 0$

$$\pi = \frac{m(1-m)}{2m(1-m)} = \frac{1}{2},$$

which verifies the earlier statement that there needs to be association in order to detect linkage.

Spielman et al. (1993) state that although the calculations were carried out using a recessive mode of inheritance for the disease, "it is easy to show that expression (3.1) provides an appropriate $\chi^2$ test of linkage whatever the penetrance values and ascertainment procedures, implying that in all cases only heterozygous ($M_1M_2$) parents should be used in the test."(Spielman, McGinnis, and Ewens 1993) This is a

result of the fact that no matter what the mode of inheritance, heterozygous parents provide no information on linkage for this test statistic.

It was stated earlier that the TDT is a test for linkage when there is known to be association, however the TDT can be used as a test for linkage and association. In this case, the hypothesis are

$H_O$: No linkage and/or no association between the marker and disease locus.
$H_A$ : Linkage and association between the marker and disease loci.

or in terms of $\theta$ and $\delta$,

$H_O$: $\theta = 1/2$, $\delta = 0$ or both
$H_A$: $\delta \neq 0$ and $\theta < 1/2$.

It should be noticed that the use of this test statistic assumes that there is no segregation distortion. **Segregation distortion** occurs when heterozygotes preferentially transmit one allele over another. If there is segregation distortion, this means that the probability of the heterozygote passing a particular allele is not be 1/2, regardless of the disease status of the offspring. If the possibility of segregation distortion exists, the transmission of alleles from heterozygous parents to affected offspring could be compared to the transmission of alleles from heterozygous parents to unaffected offspring. If these transmission probabilities are not different there is no segregation distortion.

## 3.1 Power of the TDT

The power of the TDT depends on $\pi$ which depends on $m$, $p$, $\delta$, $\theta$ and the mode of inheritance. It should be noted that $\delta$ is limited by the values of $m$ and $p$. Since $\delta = P(M_1D_1) - mp$, $-mp \leq \delta \leq 1 - mp$. The power for the TDTis evaluated under the assumption of a recessive disease.

For each value of $n$ (the number of heterozygous parents), we can use the binomial distribution with probability 1/2 to calculate the values Q and W, such that if the number of parents who transmit $M_1$ and do not transmit $M_2$, $b_{obs}$, is less than or equal to Q, or greater than or equal to W, the null hypothesis is rejected. A difference from $\pi = 1/2$ in either direction gives evidence for linkage, since there may be positive or negative association.

Then, the power of this binomial test is

$$P[(b_{obs} \leq Q) \cup (b_{obs} \geq W)] = P(b_{obs} \leq Q) + P(b_{obs} \geq W)$$

for given values of $m$, $p$, $\delta$ and $\theta$. For the following figures, the normal approximation to the binomial was used to smooth the power curves. Because of the discrete nature of the binomial, the power can vary significantly for small values of $n$ which makes unsmoothed power curves difficult to examine.

As can be seen in Figure 3.1 ($\delta$ denoted as "delta" and $\theta$ denoted as "theta") the power increases as $\delta$ increases and $\theta$ decreases. This is expected as increased association would lead to greater preferential transmission.

Figure 3.2 shows no obvious pattern for the effect of the allele frequencies on the power of the TDT although it is obvious that the power does depend on $m$ and $p$. Inspection of (3.4) shows that (in general) power should increase as $p$ decreases and the power should decrease as $m$ increases, which careful inspection of Figure 3.2 confirms.

## 3.2 An Example of the TDT Test

For an example of the TDT test, consider the data from Spielman et al (1989). The data consisted of alleles at a marker (the tandem-repeat DNA, 5' flanking polymorphism [5'FP], adjacent to the insulin gene on chromosome 11p) for 94 families with at least two children affected with insulin dependent diabetes mellitus (IDDM). There were three alleles present so alleles were classified as class 1 or class X (which consists of class 2 and class 3 alleles), due to the fact that previous studies had shown

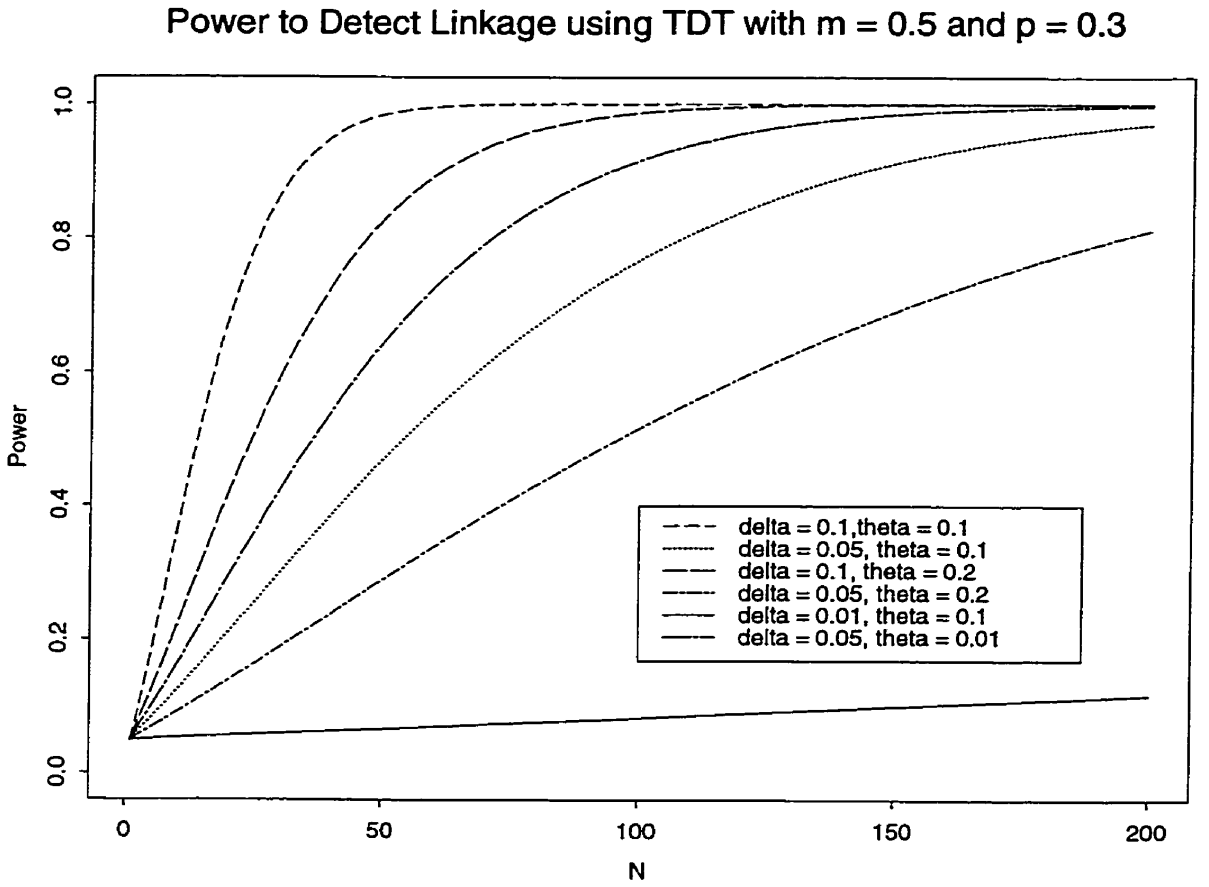


Figure 3.1: Power to Detect Linkage using TDT with $m = 0.5$ and $p = 0.3$

association between the class 1 allele and IDDM. The alleles were determined based on restriction fragment length, which is one method to determine the alleles that are present at a locus.

There were 53 families that had at least one parent heterozygous for the class 1 allele and this yielded 124 heterozygous parent-affected offspring pairs. The data is given in Table 3.4.

The test statistic is

$$T_{obs} = \frac{(b-c)^2}{(b+c)}$$

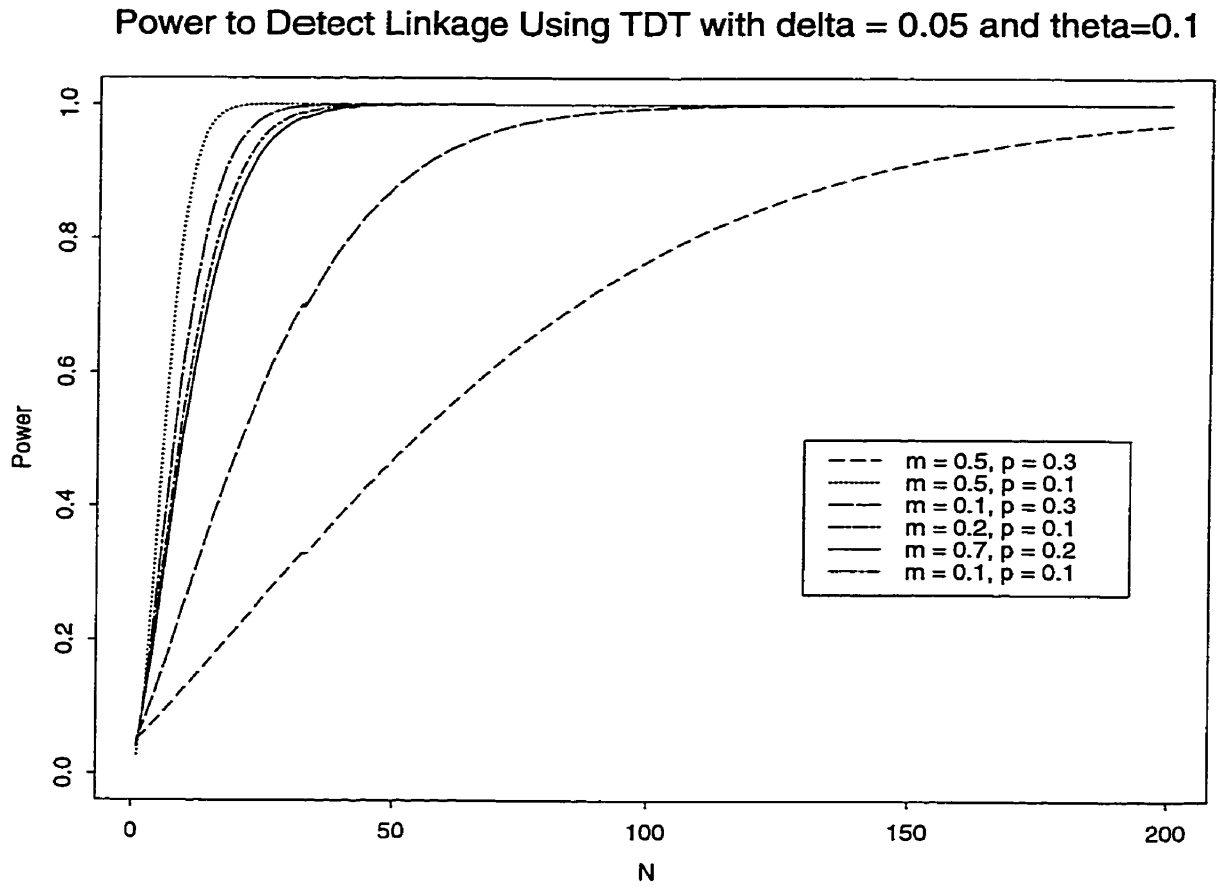


Figure 3.2: Power to Detect Linkage using TDT with $\delta = 0.05$ and $\theta = 0.1$

$$= \frac{(78-46)^2}{124}$$
$$= 8.26$$

The P-value is approximately

$$P = P(\chi_1^2 > T_{obs}) = 0.004$$

Thus there is very strong evidence of both linkage and association between the 5'FP locus and insulin dependent diabetes mellitus.

| No. of Alleles Transmitted | | |
|---|---|---|
| 1 | X | Total |
| 78 | 46 | 124 |
| $b$ | $c$ | $b+c$ |

Table 3.4: Data for Example of TDT

## 3.3 The TDT With Affected Siblings

The situation in which there are two affected siblings in a family allows for a comparison of the TDT with more standard tests for linkage as well as showing how the TDT can be extended to families with more than one affected child.

Consider the situation in which there are $h$ heterozygous parents in the sample of affected sib pairs. It is possible to divide the parents into three categories based on the alleles that they transmit to their affected offspring. The categories are parents who transmit $M_1$ to both children, parents who transmit $M_2$ to both children and parents who transmit $M_1$ to one child and $M_2$ to the other. The number of parents in each category are $i$, $j$ and $h-i-j$ respectively.

As mentioned earlier, the only values in Table 3.1 needed for the TDT are $b$ and $c$. It is possible to write these in terms of $h$, $i$ and $j$,

$$\begin{aligned} b &= 2i + (h-i-j) = h+i-j \\ c &= 2j + (h-i-j) = h-i+j. \end{aligned} \tag{3.5}$$

This makes it possible to write $b-c = 2i-2j$ and $b+c = 2h$ so the TDT statistic (3.1) can be written as

$$\chi^2_{td} = \frac{2(i-j)^2}{h}. \tag{3.6}$$

It is informative to compare this test to the "mean haplotype sharing" test proposed by Blackwelder and Elston (1985), which is the $\chi^2$ version of the ASP test presented earlier as test (2.3) which looks at the number of allele pairs that are identical by descent. Assuming that all the parents are heterozygous, there are $n = h/2$

sib pairs, and $n_1 + 2n_2 = i + j$ allele pairs IBD. This is a result of the fact that if a heterozygous parent transmits the same allele to both offspring, the alleles in the offspring are IBD. The test statistic (2.3) ($\chi^2_{hs} = T_2^2$), is

$$\begin{aligned} \chi^2_{hs} &= \frac{(n_1 + 2n_2 - n)^2}{n/2} \\ &= \frac{[i + j - (h/2)]^2}{h/4} \\ &= \frac{(2i + 2j - h)^2}{h}. \end{aligned} \tag{3.7}$$

There is a relationship between $\chi^2_{hs}$ and $\chi^2_{td}$ that is not readily apparent. Although both tests are valid tests for linkage between the marker and disease locus, they test for it in different ways, as the mean haplotype sharing test does not take into account the alleles that are at the marker locus.

The three categories that the heterozygous parents were divided into ($i$, $j$ and $h - i - j$) have probabilities of 1/4, 1/4 and 1/2 under the null hypothesis of no linkage. It is possible to do a goodness of fit test for these categories, and the "total" test statistic is

$$\chi^2_{total} = \frac{(i - h/4)^2}{h/4} + \frac{(h - i - j - h/2)^2}{h/2} + \frac{(j - h/4)^2}{h/4}. \tag{3.8}$$

However, this can be written as

$$\begin{aligned} \chi^2_{total} &= \frac{(i - h/4)^2}{h/4} + \frac{(h - i - j - h/2)^2}{h/2} + \frac{(j - h/4)^2}{h/4} \\ &= \frac{4(i^2 - hi/2 + h^2/16) + 2(h^2/4 - hi - hj + 2ij + i^2 + j^2) + 4(j^2 - hj/2 + h^2/16)}{h} \\ &= \frac{(2i^2 - 4ij + 2j^2) + (4i^2 + 8ij - 4hi - 4hj + 4j^2 + h^2)}{h} \\ &= \frac{2(i - j)^2}{h} + \frac{(2i + 2j - h)^2}{h} \\ &= \chi^2_{td} + \chi^2_{hs}. \end{aligned}$$

Since $\chi^2_{total}$ can be written as the sum of $\chi^2_{hs}$ and $\chi^2_{td}$, $\chi^2_{hs}$ and $\chi^2_{td}$ use the data in two statistically independent ways to test the hypothesis that the disease and marker loci are linked.

## 3.4 Example of TDT with Affected sibs

The data that was used in Section 3.2 can also be used as an example for the TDT with affected sibs. To show the power of the TDT in comparison to standard linkage tests, the TDT with affected sibs is compared to the mean haplotype sharing test of Blackwelder and Elston (1985) using the same data set. There were 45 families that met both the criterion for the TDT (at least one parent heterozygous) and for the haplotype sharing test (informative sib pairs). For the informative families that had two heterozygous parents, only one heterozygous parent is used to ensure the fairness of the comparison.

The data, divided into the three classes as in Section 3.3, are shown in Table 3.4.

| No. of Parents who Transmit | | | |
|---|---|---|---|
| Class 1 to Both Children | Class 1 to One Child Class X to Other | Class X to Both Children | TOTAL |
| $i = 15$ | $h - i - j = 24$ | $j = 6$ | $h = 45$ |

Table 3.5: Transmission of Alleles from 45 I/X Parents of Affected Sib Pairs

The TDT statistic (3.6) is

$$\begin{aligned} \chi^2_{td} &= \frac{2(i-j)^2}{h} \\ &= \frac{2(15-6)^2}{45} \\ &= 3.6 \end{aligned}$$

giving a P-value of $P(\chi^2_1 > 3.6) = 0.058$, which is borderline significant by the statistical standard of $\alpha = 0.05$. The haplotype sharing test statistic (3.7) is

$$\begin{aligned} \chi^2_{hs} &= \frac{(2i + 2j - h)^2}{h} \\ &= \frac{[2(15) + 2(6) - 45]^2}{45} \\ &= 0.2, \end{aligned}$$

giving a P-value of $P(\chi^2_1 > 0.2) = 0.65$ which is not significant, nor close to being significant. One thing that should be noted is that there were actually fewer allele pairs IBD ($i+j = 21$) than would be expected under the null hypothesis of no linkage ($h/2 = 22.5$). Thus, the difference being detected by the haplotype sharing test is actually in the opposite direction of what would be expected if there was linkage.

This data shows the power of the TDT to use a known association to detect linkage that is not detectable by use of affected sib pairs.

# Chapter 4

# Extensions of Tests and Summary of Thesis

Since the various papers discussed in the previous chapters were published there have been several extensions to the test procedure. There have been several extensions to linkage tests as well as the TDT.

## 4.1 Extensions to Linkage Tests

Most extensions of linkage tests try to deal with the difficulty of obtaining affected relative pairs.

One minor extension to the many tests for linkage is genotype reconstruction. **Direct reconstruction** consists of using other relatives to reconstruct the genotypes of individuals who are not available for genotyping to allow for unambiguous determination of IBD status for affected relatives.

Identity by state methods can also be used to detect linkage in affected relative pairs that can not have their IBD status determined.

Discordant relative pairs (one affected by disease, one unaffected) can also be used to detect linkage although these tests have significantly less power that affected relative pair tests.

### 4.1.1 Direct Genotypic Reconstruction

As previously mentioned, it can be difficult to determine the IBD status of affected relative pairs, but the use of other relatives can allow determination of the IBD status of affected relative pairs that could not have been determined by the usual criteria.

Consider the situation in which we have affected sibs whose parents are unavailable for typing. It had been previously stated that it is necessary to be able to type the parents for these siblings to be informative, but in some cases it is not necessary. If there are other relatives available, such as other siblings or grandparents of the affected sibs, it may be possible to determine what alleles the parents had and thus it may be possible to determine the IBD status of the affected siblings.

Consider the situation in which two affected siblings with genotypes of AB and AC at the marker locus have parents who are unavailable to be genotyped at the marker locus. If these affected siblings have another sib with genotype BC at the marker locus, then we can determine that the B and C alleles came from different parents. Thus, the genotypes for the parents are AB and AC, and the affected sibs have no alleles IBD, because the A alleles that they have in common had to come from separate parents. This approach can be especially useful with late-onset diseases in which the parents are often unavailable for genotyping.

Direct genotypic reconstruction can also be used when one of the affected relatives is unavailable for genotyping if that individual had offspring. The idea is identical to the example given above, in which the offspring are used to reconstruct the genotype of the parent.

There are many other situations in which this method can be attempted, but the use of this method will not always determine the IBD status, because even with the full genotypes of all people involved, the relative pair may still be uninformative.

### 4.1.2 The Identity by State Method of Linkage Analysis

Since it is often difficult to obtain informative affected sib pairs, Lange (1986) proposed a test for linkage based on identity by state (IBS) rather than IBD. Identity

by state focuses on the alleles that a sib pair have present rather than on the transmission of alleles. Two sibs are said to be marker concordant if they have the same marker genotypes and marker discordant if they share no alleles in common at the marker locus. If they have one allele in common at the marker locus, they are said to be half-concordant. For extremely polymorphic loci (many alleles), the probabilities for being marker concordant, half concordant and discordant approach 1/4, 1/2 and 1/4 respectively, as marker concordant is almost equal to having two alleles IBD. This is based on the fact that if the marker locus is highly polymorphic, the parents are not likely to be homozygous, nor are they likely to have the same alleles (Lange 1986). The idea of the test is similar to that of IBD tests, comparing the number of observed concordant or discordant sib pairs with the number expected, but it is complicated to calculate the expected counts, as they are dependent on the number of alleles present and the allelic probabilities.

Identity by state methods make it easier to obtain subjects, as any affected relative pair can be used, but they can often give type I errors if incorrect values for the number of alleles and allelic proportions are used to obtain the expected values and variance. Also, it is difficult to give a general test statistic for these tests as both the expected values and the variance are greatly dependent on the polymorphism at the marker locus. For this reason, the identity by state method seems unreliable and is rarely used.

### 4.1.3 Discordant Relative Pairs

Discordant relative pairs have one affected and one unaffected individual. Other than that difference, the tests for linkage based on IBD status are identical in form to test using affected relative pairs. The tests now look for fewer alleles IBD than are expected under the null hypothesis of no linkage.

Risch (Risch 1990b) calculated the probabilities of a discordant relative pair having $i$ alleles IBD, $y_{Ri}$, using a similar method as for affected relative pairs. Ignoring the possibility of recombination,

$$\begin{aligned} y_{R0} &= P(IBD = 0|1 \text{ relative affected, } 1 \text{ affected}) \\ &= \frac{P(IBD = 0)P(1 \text{ relative affected, } 1 \text{ affected}|\text{IBD} = 0)}{P(1 \text{ relative affected, } 1 \text{ affected})} \\ &= \alpha_{R0}\frac{K(1-K)}{1-K_R} \\ &= \alpha_{R0}\frac{1-K}{1-K_R}. \end{aligned}$$

Similarly,

$$y_{R1} = \alpha_{R1}\frac{1-K_O}{1-K_R}$$

and

$$y_{R2} = \alpha_{R1}\frac{1-K_M}{1-K_R}.$$

Risch (1990b) showed that the deviations from the null proportions, $\epsilon_{Ri} = y_{Ri} - \alpha_{Ri}$, are

$$\epsilon_{R0} = \alpha_{R0}\frac{K}{1-K_R}(\lambda_R - 1),$$

$$\epsilon_{R1} = -\alpha_{R1}\frac{K}{1-K_R}(\lambda_O - \lambda_R),$$

and

$$\epsilon_{R2} = -\alpha_{R2}\frac{K}{1-K_R}(\lambda_M - \lambda_R).$$

Using equations (2.10)-(2.12), the deviations, $\delta_{Ri} = z_{Ri} - \alpha_{Ri}$, from the null probabilities for affected relatives that share $i$ alleles IBD are

$$\delta_{R0} = -\frac{\alpha_{R0}}{\lambda_R}(\lambda_R - 1),$$

$$\delta_{R1} = \frac{\alpha_{R1}}{\lambda_R}(\lambda_O - \lambda_R)$$

and

$$\delta_{R2} = \frac{\alpha_{R2}}{\lambda_R}(\lambda_M - \lambda_R).$$

From these, $\epsilon_{Ri} = -[K_R/(1-K_R)]\delta_{Ri}$. As mentioned, the deviation from the null probabilities for discordant relative pairs is in the direction opposite to that of affected relative pairs. $\epsilon_{Ri}$ will be less than $\delta_{Ri}$ if the reccurence risk $K_R < 0.5$. Risch (Risch 1990b) states that "values for $K_R$ are always less than 50% and usually less than 25% (except for high-penetrance autosomal dominant diseases)", so the deviations are smaller for discordant relative pairs and thus discordant relative pair tests have less power than affected relative pair tests.

Whether it is better to use linkage tests based on affected relative pairs or discordant relative pairs depends on the availability for subjects of the two tests. Discordant relative pairs provide significantly less power, but subjects are usually easier to obtain.

## 4.2 Extensions to the TDT

The TDT test has become very popular among geneticists and genetic epidemiologists and there have been many papers published about it and many minor modifications to the test. Although originally designed to be a test for linkage in the presence of association, it has been used as a test for association in the presence of linkage (Martin, Kaplan, and Weir 1997). It has been modified into the "Sib-TDT" which allows the use of unaffected sibs (Spielman and Ewens 1998) and the Pedigree Disequilibrium Test (PDT) which allows use of extended pedigrees (Martin, Monks, Warren, and Kaplan 2000).

## 4.3 The Future of Genetic Studies of Complex Human Diseases

In a paper entitled "The Future of Genetic Studies of Complex Human Diseases" Risch and Merikangas (1996) compared affected relative pair tests and the TDT. They showed that the power of the TDT is substantially greater than that of the ASP tests, especially for diseases with small relative risk ratios. In some of their calculations the number of affected relative pairs needed to obtain 80% power was over 200 times greater than the number of heterozygous parent-affected offspring pairs needed to obtain the same power with the same assumptions.

Risch and Merikangas (1996) state that the only limit to using the TDT test in a genome-wide search for loci or alleles that influence a disease is a technological one, not a statistical one. To be able to use the TDT to detect genes with a small effect that would not be detectable with linkage tests, Risch and Merikangas state that "a larger number of genes (up to 100,000) and polymorphisms (preferentially ones that create alterations in derived proteins or their expression) must first be identified" (Risch and Merikangas 1996).

Risch and Merikangas (1996) finish their paper with a quote worth repeating in its entirety:

> The human genome project can have more than one reward. In addition to sequencing the entire human genome, it can lead to identification of polymorphisms for all the genes in the human genome and the diseases to which they contribute. It is a charge to the molecular technologists to develop the tools to meet this challenge and provide the information necessary to identify the genetic basis of complex human disease (Risch and Merikangas 1996).

## 4.4 Summary of Thesis

The test for disease-allele association is a good start when trying to determine what part of the genome affects a certain disease, as it is a simple test that does not require a lot of information nor does it make any major assumptions. However, false positive results can occur as a result of population stratification, so the results from a test of association are of limited use.

Linkage tests are not subject to any problems of population stratification but have poor power and it may be difficult to get adequate sample sizes to detect linkage if it exists.

The TDT has better power than linkage tests and is not affected by population stratification but there has to be linkage disequilibrium in order to detect linkage, thus it may not detect linkage when linkage is present.

Despite the problem of association being needed to detect linkage with the TDT, the TDT seems to be the best test to use in order to determine what parts of the genome have an influence on a given disease, as it offers the best power and it is easier to obtain subjects. The main problem that prevents the test from being used is the lack of marker loci spaced out over the whole genome. One hopes that this is a problem that can be overcome.

# Bibliography


Bishop, T. D. and J. A. Williamson (1990). The power of identity-by-state methods of linkage anaylsis. *American Journal of Human Genetics* **46**, 254–265.

Blackwelder, W. and R. Elston (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. *Genetic Epidemiology* **2**, 85–97.

Boehnke, M. and C. Langefeld (1998). Genetic association mapping based on discordant sib pairs: The discordant alleles test. *American Journal of Human Genetics* **62**, 950–961.

Haseman, J. and R. C. Elston (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavioural Genetics* **2**, 3–19.

Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics* **52**, 362–374.

James, J. (1971). Frequency in relatives of an all-or-none trait. *Annals of Human Genetics* **25**, 47–49.

Khoury, M., T. Beaty, and B. Cohen (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press.

Knapp, M. (1997). *Genetic Mapping of Disease Genes*, Chapter The affected sib pair method for linkage analysis, pp. 147–157. Academic Press Ltd.

Lange, K. (1986). A test statistic for the affected-sib-set method. *Annals of Human Genetics* **50**, 283–290.

Lynch, M. and B. Walsh (1998). *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Inc.

Martin, E., N. Kaplan, and B. Weir (1997). Tests for linkage and association in nuclear families. *American Journal of Human Genetics* **61**, 439–448.

Martin, E., S. Monks, L. Warren, and N. Kaplan (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *American Journal of Human Genetics* **67**, 146–154.

Motro, U. and G. Thomson (1985). The affected sib method. i. statistical features of the affected sib-pair method. *Genetics* **110**, 525–538.

Ott, J. (1989). Statistical properties of the haplotype relative risk. *Genetic Epidemiology* **6**, 127–130.

Risch, N. (1987). Assessing the role of hla-linked and unlinked determinants of disease. *American Journal of Human Genetics* **40**, 1–14.

Risch, N. (1990a). Linkage strategies for genetically complex traits. i. *American Journal of Human Genetics* **46**, 222–228.

Risch, N. (1990b). Linkage strategies for genetically complex traits. ii the power of affected relative pairs. *American Journal of Human Genetics* **46**, 229–241.

Risch, N. (1990c). Linkage strategies for genetically complex traits. iii the effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics* **46**, 242–253.

Risch, N. and K. Merikangas (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Smith, C., D. Falconer, and L. Duncan (1972). A statistical and genetical study of diabetes. *Annals of Human Genetics* **35**, 281–299.

Spielman, R., M. Baur, and C.-D. F (1989). Genetic analysis of iddm: summary of gaw5-iddm results. *Genetic Epidemiology* **6**, 43–58.

Spielman, R. and W. Ewens (1998). A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**, 450–458.

Spielman, R., R. McGinnis, and W. Ewens (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). *American Journal of Human Genetics* **52**, 506–516.

Suarez, B. K. (1983). A sib-pair strategy for the use of restricted fragment length polymorphisms to study the mode of transmission of type ii diabetes. *American Journal of Human Genetics* **35**, 34–48.

Walker, A. and A. Cudworth (1980). Type 1 (insulin dependent) dibetes multiplex families. *Diabetes* **29**, 1036–1039.

Weeks, D. and K. Lange (1988). The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics* **42**, 315–326.

Whittemore, A. S. and I.-P. Tu (1998). Simple, robust linkage tests for affected sibs. *American Journal of Human Genetics* **62**, 1228–1242.