Human genetic susceptibility to leprosy: methodological issues in a linkage analysis of extended pedigrees from Karonga district, Malawi

Thesis submitted for the degree of Doctor of Philosophy

Catriona (Chris) Wallace

London School of Hygiene and Tropical Medicine

Faculty of Medicine,
University of London

2003

ABSTRACT

Leprosy is a disease of humans which is caused by infection with *Mycobacterium leprae*. Although infection with *M. leprae* is necessary for disease, it is thought only a small proportion of those infected (probably less than 10%) develop clinical disease, which may be manifested across a wide spectrum. Susceptibility to leprosy is influenced by both genetic and non-genetic factors, and there is evidence that genetic influences vary between populations.

Linkage analysis is a method for finding genes that influence a particular trait. Nonparametric methods of linkage analysis compare the identity by descent (IBD) sharing of genes among affected relatives to that expected in the absence of linkage. Often nonparametric linkage analysis is applied to small families with multiple affected siblings, but extended multicase pedigrees may offer increased power to detect genetic determinants of disease. This thesis deals with methodological issues raised in a linkage analysis of extended pedigrees with multiple cases of leprosy.

Power to detect linkage using affected relative pair data can be expressed as a function of an epidemiological parameter called the relative recurrence risk ratio (λ_R) which is defined as the ratio of the risk of disease in particular relatives of cases to the population risk of disease. It will be shown that power to detect linkage using relative trios can also be expressed as a function λ_R and a second, related parameter, $\lambda_{R,R}$, defined here as the ratio of the risk of disease in individuals who have two affected relatives to the population risk of disease. Estimates of λ_R can be inflated if environmental risk factors are not properly accounted for. Methods for estimating λ_R and λ_{RR} while accounting for environmental risk factors are investigated using a marginal model, and estimates of λ_R are presented for first, second and third degree relatives of people affected by leprosy in Karonga district, Malawi.

Further work examines the selection of members from extended pedigrees for inclusion in a linkage analysis. Simulation techniques are used to select members in order to maximise information about IBD sharing between affected pedigree members without typing unnecessary members, which can be time-consuming and costly. Finally, nuclear families and extended pedigrees from Karonga are used in a partial genome screen to search for chromosomal regions which may be linked to susceptibility either to leprosy per se or to leprosy type. Preliminary results are presented: no regions show significant evidence of linkage according to the stringent criteria applied in genome screens, but there are regions that show evidence for potential linkage that would be of interest to follow-up in this population.

CONTENTS

Appendices 237

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

γ	Genotype relative risk
γ_1	Fisher skewness
λ_R	Relative recurrence risk ratio
λ_S	Sibling recurrence risk ratio
ASP	Affected sibling pair
BCG	Bacille Calmette-Guerin
cM	centi-Morgan
CRM	Cross ratio model
HLA	Human Leukocyte Antigen
HS	Half sibling
IBD	Identity by descent
IBS	Identity by state
KPS	Karonga Prevention Study
MB	Multibacillary
MHC	Major Histocompatibility Complex
MLB	Maximum likelihood binomial
MLS	Maximum lod score
NDV	No dominance variance
PB	Paucibacillary
TB	Tuberculosis
TNF	Tumor Necrosis Factor
VDR	Vitamin D receptor

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisors, Paul Fine and David Clayton for making this thesis possible, for the time they have given me and their invaluable advice. I have learnt a great deal from my discussions with them.

I would like to thank the Medical Research Council for their financial support during this project. The data analysed here were collected within the context of the Karonga Prevention Study (KPS), which has been supported by the British Leprosy Relief Association (LEPRA) and the Wellcome Trust, with contributions from the International Federation of Anti-Leprosy Organisations (ILEP) and the WHO/UNDP/World Bank Special Programme for Research and Training in Tropical Diseases.

I wish to acknowledge the special contributions of Drs Jorg Ponnighaus, David Warndorff and Mia Crampin, and more than fifty field staff for the fieldwork; Lyn Bliss and Keith Branson for their data management; the people of Karonga for their participation in these studies and the Health Sciences Research Committee of Malawi for encouraging the KPS programme.

Professor Adrian Hill and Drs Jodene Fitness and Branwen Hennig have shown great enthusiasm for this project. In particular, Drs Fitness and Hennig carried out the majority of the genotyping on which these results are based and I would like to thank them and their colleague Drs Christophe Aucun for their work in this. I would also like to thank them and other members of the Hill group including Drs Graham Cooke and Ruby Siddiqui for helpful discussions.

I thank Prof. Laurent Abel for kindly making available a test version of software.

I also wish to thank the many friends I have made at the London School of Hygiene and Tropical Medicine for their support and for successfully distracting me from my work on many occasions during the last three years. Particular thanks are due to Cynthia Braga, Najibullah Habib, Ali-Akbar

Haghdoost, Sukhum Jiamton, Sue Lee, Maria-Claudia Nascimento, Anita Ramesh, Brenda Roche, Sunheang Shin, Eric Tongren, Salim Vohra and Leland Yee. I hope our friendship will continue for many years.

Last but by no means least, to my partner Jan, I offer my love and gratitude for patient support and understanding during the last three years.

DEDICATION

I dedicate this thesis to my parents, who I hope would have been proud.

CHAPTER 1.
INTRODUCTION

Leprosy is a disease caused by infection with *Mycobacterium leprae*. Infection leads to disease in only a minority of those infected and, together with non-genetic factors, host genetics are thought to influence susceptibility to disease. The Leprosy Evaluation Project (LEP) began in 1978 in Karonga District, Northern Malawi, as a prospective study of the prevalence and incidence rates of clinical leprosy. It further aimed to investigate risk factors which may be associated with both infection and disease. Among other work, the project, now known as the Karonga Prevention Study (KPS), has conducted two total population surveys, collecting data on $\sim$250,000 individuals, several of whom are members of large pedigrees containing multiple individuals affected by leprosy. This project arose with the aim of investigating this unique dataset in order to quantify the role of host genetics in affecting susceptibility to leprosy and to examine whether positive results from genetic studies of leprosy in other populations could be replicated in this population.

In this chapter, the specific aims and objectives of this project are introduced and the structure of the rest of this thesis is described.

1.1 Aims and objectives

There are three main aims of this project:

1. To quantify the risk of disease to relatives of affected individuals. This is a measure of the size of genetic effect and is achieved by estimating the relative recurrence risk ratio (λ_R) among first, second and third degree relatives from Karonga district, Malawi. This aim led to the following specific objectives
    (a) To quantify the effect of non-genetic factors in this population.

    (b) To construct a model that would allow the residual risk of disease in relatives of affected individuals to be estimated, after accounting for the effect of these non-genetic factors.
    (c) To implement the model in statistical software and apply it to data from the KPS.

2. To examine how extended multicase pedigrees available from Karonga could be efficiently used in a linkage analysis. This was done by considering the choice of affected and unaffected members separately and led to the objectives
    (a) To explore the aggregation of disease among relative pairs and trios in order to make inference about the use of affected relative trios and above in family-based genetic studies.
    (b) To examine sampling strategies for unaffected members of multiplex pedigrees using simulation. Comparisons were made on the basis of the expected information about IBD sharing that would be obtained under different strategies.

3. To examine whether specific results in other studies of genetic susceptibility to leprosy could be replicated in the Karonga population. Extended multicase pedigrees were available, and the specific objectives in this part of the study were
    (a) To conduct a partial genome screen of selected chromosome regions identified as showing association or linkage to leprosy in other genetic studies to search for areas that showed evidence of linkage to leprosy per se.
    (b) To conduct a partial genome screen of the same regions to search for areas that showed evidence of linkage to clinical type.

1.2 Structure of this document

This document describes the work done to meet the above aims. Chapter 2 introduces the genetic terminology that will be used in this document, and, in particular, defines and describes the way in which genetic linkage arises. Statistical methods for detecting linkage are reviewed. There then follows a review of the basic epidemiology of leprosy and the factors that are known to influence a person's susceptibility to disease. In particular, there is a detailed review of the evidence that genetic factors have a role to play and a description of the areas of genome that have been implicated.

Chapter 3 provides an introduction to the Karonga Prevention Study (KPS) and the history of leprosy in Karonga. There is a descriptive analysis of the epidemiological variables which will be used in this project and a preliminary analysis of their effect on disease susceptibility in this population.

In chapter 4, a method is developed to estimate the relative recurrence risk, λ_R, while accounting for non-genetic risk factors, and estimates obtained from fitting this marginal model to the KPS data are presented. Extension of this model to relative trios is also explored.

Chapter 5 describes how the related second degree recurrence risk ratio $\lambda_{R,R}$ varies under different one-locus genetic models and how power to detect linkage using relative trios and pairs may be estimated under these models. Results of the power calculations are presented for half sibling pairs and trios.

The linkage analysis is discussed in chapter 6. First, the choice of strategy for the partial genome screen and the statistical methods used are discussed. Efficient use of the extended pedigrees is explored using simulation and recommendations are made for other studies of extended pedigrees. Finally, the results of the analysis are presented and related to those from other studies.

Lastly, chapter 7 describes how the above aims and objectives have been met and strength of evidence for genetic influence of susceptibility to leprosy is reviewed in light of the results of this project.

CHAPTER 2.
BACKGROUND

2.1 Introduction

This chapter describes the background to this project. Some genetic terminology and parameters are introduced in sections 2.2 and 2.3. Methods for linkage analysis of binary traits are reviewed in section 2.4 and linkage analysis studies of infectious diseases are discussed in section 2.5.

Sections 2.6–2.7 introduce leprosy, describing the natural history, non-genetic risk factors for disease and the evidence for host genetic risk factors.

2.2 Basic genetic concepts

In this section, some terminology commonly used in genetics is introduced and the manner in which genetic material is passed from one generation to the next and how linkage arises is described.

2.2.1 Terminology and definitions

Genes are coded by strings of deoxyribonucleic acid (DNA) molecules. Within a given gene, *exons* are the portions that code for proteins and *introns* are the sequences between exons, that are not transcribed. The physical site or location of a gene is called a *locus*, and the different forms of a gene that may exist are called *alleles*. Genes are arranged in chromosomes; in humans, there are 46 chromosomes arranged in pairs, with one member of each pair of chromosomes inherited from a person's father, the other from their mother. There are 22 *autosomal* pairs (numbered 1 to 22) and one pair of sex chromosomes, which determine a person's sex. Females have two X chromosomes while males have one X (inherited from their mother) and one Y (inherited from their father).

Each chromosome has particular features. The *centromere*, though not the actual centre of the chromosome, separates the short (p) and long (q) arms from one another and *telomeres* are located at either end.

2.2.2 Meiosis and recombination

The process by which genetic material is passed from parents to a child is called meiosis; a schematic diagram of this process is shown in figure 2.1. During human meiosis, two copies of each chromosome line up and exchange genetic material by means of crossovers. When there is an even number of crossovers (including zero) between two loci, then the genes at those loci will be present together in the resulting gamete. An odd number of crossovers will result in the genes being separated in the resulting gametes. Genes at two loci are said to be *recombinant* if they are not passed on together and *non-recombinant* otherwise.

2.2.3 Map distance and recombination fractions

The map distance x between two loci, measured in Morgan units (M), is defined as the expected number of crossovers between them. More commonly, the distance is expressed in centi-Morgans (cM), with 100cM = 1M.

The recombination fraction, θ, is the probability that genes at two loci are recombinant - ie the probability there are an odd number of crossovers between them. Note that we expect $\theta = \frac{1}{2}$ for genes at loci on separate chromosomes, since they are equally likely to be recombinant or non-recombinant. However, crossover events between two loci are increasingly rare for loci close together on the same chromosome, and so $\theta \rightarrow 0$ as we consider loci closer and closer together.

Conversion between map distance and recombination fraction requires that assumptions are made about how crossovers arise, and several *map functions* have been proposed. The map distance is the expected number of crossovers between two loci in a single chromatid, whereas the recombination fraction is the probability there is an odd number of crossovers between the two. Mather deduced the recombination fraction was 1/2 as long as there was at least one crossover between two loci. If p_0 is the probability of no crossovers, then the recombination fraction is

$$\theta = (1 - p_0) \times 1/2 + p_0 \times 0 = \frac{1 - p_0}{2}.$$

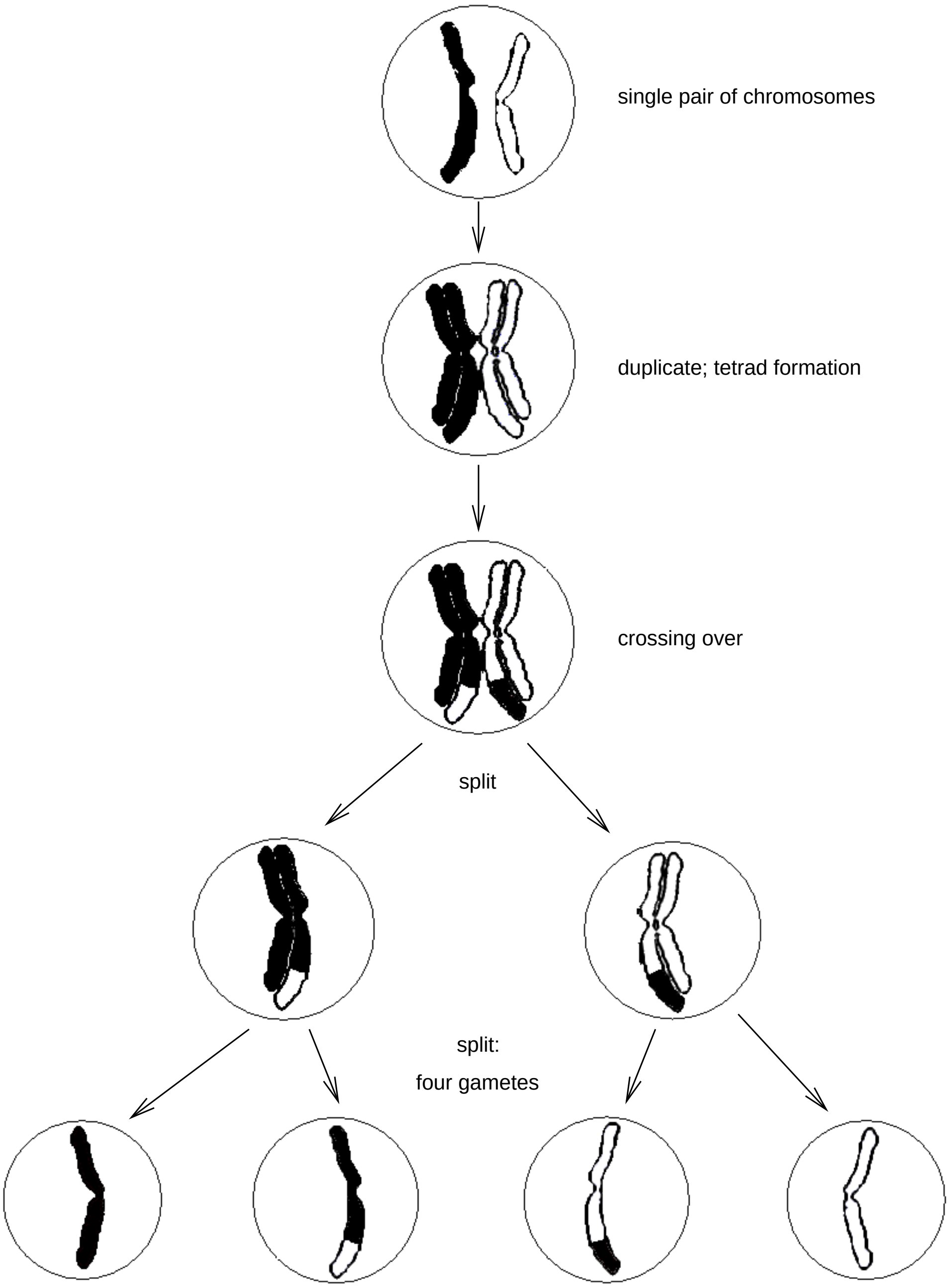


Figure 2.1: Schematic diagram of meiosis. Each of the autosomal chromosomes in the parent cells duplicate and line up. They exchange material by crossing-over at places called chiasmata*. They then split to form four gamete cells (eggs in women, sperm in men). The haploid chromosomes in gametes from each parent will pair up in the child cell, so that the child has two copies of each chromosome, one from each parent*

The simplest map function, known as the Morgan map function, assumes chromosomal segments can have at most one crossover, and that the probability of a crossover is proportional to the length of the segment. The probability of a crossover in m map units is therefore $2m$ and

$$\theta = \frac{1 - p_0}{2} = \frac{1 - (1 - 2m)}{2} = m.$$

Note that the function is only valid for $m < 1/2$ (otherwise we would have $\theta > 1/2$) so is not applicable for long segments of chromosome. The Haldane map function assumes crossovers occur at random, independently of one another. This implies a Poisson process, with rate $2m$ in a segment of length m. So $p_0 = e^{-2m}$ and

$$\theta = \frac{1 - e^{-2m}}{2}.$$

However, observations show that crossovers do not tend to occur completely independently. Due to *interference*, the probability of having two crossovers very close to each other is less than that predicted by the Haldane function, and more complex arguments that take account of this lead to the Kosambi function

$$\theta = \frac{e^{4m} - 1}{2(e^{4m} + 1)}.$$

Other functions have also been proposed, but the above three are the simplest. Figure 2.2 shows that all functions produce very similar results for the small distances (< 10cM) which are generally used when conducting linkage analyses, as described in section 2.4.

2.3 The relative recurrence risk ratio, λ_R

Many diseases appear to be familial - that is, they run in families (leprosy was thought to be familial before the discovery of *Mycobacteria leprae*). The relative recurrence risk ratio (λ_R) is one measure of ‘how familial’ a disease is. A high ratio is often used as evidence for genetic influence of a trait. Note, however, that a disease may appear familial not only because of shared genetic predisposition, but also because of aggregation of non-genetic risk factors within a family, or, in the case of an infectious disease, common exposure to the infectious agent.

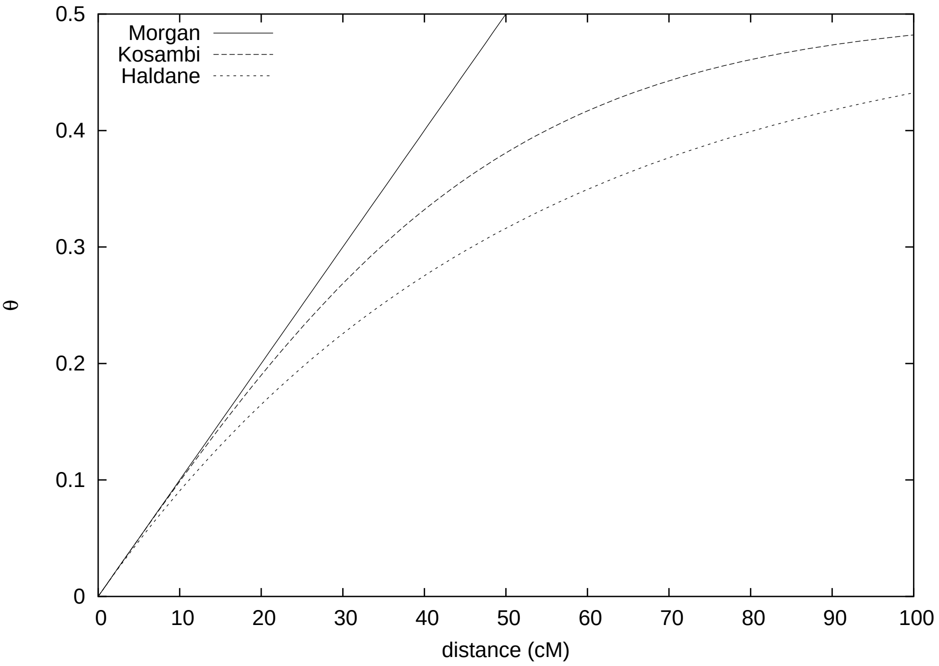


Figure 2.2: Map functions for relating map distances to recombination fractions

2.3.1 Definition

λ_R is a conceptually simple measure which has been in use for a number of years. Penrose (1953), denoting it by K, discussed interpretation under different genetic models. It is defined by Risch (1990a) as 'the risk ratio for a type R relative of an affected individual compared with population prevalence'. Mathematically, let

$$X_i = \begin{cases} 0 & \text{individual } i \text{ unaffected} \\ 1 & \text{individual } i \text{ affected} \end{cases}$$

The population prevalence is $K = E(X_1)$ and the relative recurrence risk is $K_R = E(X_2|X_1 = 1)$ where individual 2 is a type R relative of individual 1. Then define the recurrence risk ratio as

$$\lambda_R = K_R/K \tag{2.1}$$

Risch used the following result from James (1971)

$$K_R = K + \frac{\mathrm{cov}(X_1, X_2)}{K}$$

to further express

$$\lambda_R = K_R/K = 1 + (1/K^2)\,\mathrm{cov}(X_1, X_2)$$

2.3.2 Applications

Interest in λ_R, and the sibling recurrence risk, λ_S, in particular, has increased since three seminal papers by Risch (1990a,b,c). In his first paper, Risch (1990a) described the expected pattern of λ_R for different relatives under single- and multi-locus genetic models. For single locus and additive multilocus models, $(\lambda_R - 1)$ decreases by a factor of two with each degree of relationship. For multiplicative models (ie in the presence of epistasis), the decrease is steeper. Risch suggested that comparison of values for λ_R estimated from data to these predicted patterns may be used to detect epistasis. He compared published estimates of λ_R for schizophrenia for different relatives and showed it was less likely that schizophrenia was governed by a single locus than by several loci acting multiplicatively (Risch, 1990a).

If any locus is shown to affect disease, a locus-specific λ_S (or λ_R) can also be calculated representing the effect due to that locus alone. Risch (1990b) showed that locus-specific λ_R could be used to predict the power to detect linkage for different affected pairs using completely polymorphic markers. This led to recommendations about which relative pairs are likely to be more useful for different λ_R values. In practice, power is lower than that predicted by Risch because of recombination between marker and disease loci and because markers are not completely polymorphic.

λ_S is also commonly used in exclusion mapping - regions of chromosomes in a genome scan are excluded on the basis that they do not confer a λ_S of at least, say, 1.5 (e.g. Duffy et al., 2001). See section 2.4.6 for further detail on exclusion mapping.

2.3.3 Estimation

The papers by Risch created interest in λ_R, but estimation is not straightforward. Generally, λ_R has been estimated using epidemiological case-control studies in which a set of probands with disease is selected from a population registry or from a series of patients attending a clinic. Detailed family histories are taken and used to identify the number and distribution of cases among relatives. The same procedure can be applied to relatives of matched controls and the rate of disease in each group of relatives compared, allowing estimation of λ_R.

Guo (1998) has shown that ignoring or incorrectly modelling ascertainment can lead to biased estimates of λ_R and increase the the probability of falsely concluding a disease is subject to a genetic effect. Olson and Cordell (2000) further investigated ascertainment and presented methods for unbiased estimates of λ_S under different ascertainment schemes.

However, many complex diseases are influenced by environment as well as genetics, and many non-genetic factors aggregate in families. Estimates of λ_R may be inflated if environmental factors are ignored (Guo, 2000). Also, many complex diseases have onset after birth and disease is then an event in time. Individuals who do not have disease when observed at one timepoint may yet develop disease. Therefore an estimated λ_R greater than 1 may reflect either shared genetic or non-genetic factors (or both). More complicated methods of estimation which take account of such issues are discussed in section 4.3.

2.4 Genetic linkage

Linkage analysis has been used, with considerable success, in the search for disease-causing genes relating to simple Mendelian traits. It has also been used with success in the search for susceptibility genes related to more complex diseases, including three infectious diseases (Marquet et al., 1996; Bellamy et al., 2000; Siddiqui et al., 2001). These studies are discussed in detail in section 2.5.

The description of meiosis in section 2.2.2 showed how linkage may arise. If two loci lie close together on the same chromosome, they will tend to segregate together ($\theta < \frac{1}{2}$) and are said to be 'linked'. The probability of this happening is greater, and so linkage is 'tighter', the closer they are. *Linkage analysis* is concerned with estimating genetic distances between loci on a chromosome and testing whether a putative gene influencing a specific trait is linked to a marker locus or marker loci with known location.

More formally, assume there exists a locus which influences a particular trait of interest. Testing for linkage amounts to testing whether $\theta < \frac{1}{2}$ ($\theta = \frac{1}{2}$ corresponds to independent segregation of the two loci). Finding marker loci that are linked to a trait enables us to narrow the region of the genome that must be searched to find a trait-influencing locus.

Although methods exist for the analysis of both qualitative and quantitative traits, this review focuses mainly on binary traits (eg affected/unaffected) as these are of particular relevance to this study. Methods for linkage analysis of binary traits can be classified broadly into two groups: mode of inheritance based methods (often called model-based) and model-free methods (sometimes called nonparametric).

A 'complex' trait is so-called because it does not follow simple Mendelian patterns of inheritance. This can be due to a number of factors, including the trait being under the control of multiple genes; uncertain diagnosis; genetic heterogeneity (different genetic variants causing the same phenotypic trait); variable age of onset; or environmental factors (leprosy fulfils many of these criteria). Model-based methods assume the mode of inheritance of a trait is known and are more powerful than model-free because the model provides additional information, but may not be appropriate for complex traits. Misspecification of parameters can introduce bias in estimates of the recombination fraction in two point analysis (Clerget-Darpoux et al., 1986). Where the true model is unknown, it is possible to perform analyses under several different models, and then take an average over all analyses, with significance levels adjusted for multiple testing (Greenberg et al., 1998; Abreu et al., 1999). More often, model-free methods are used for complex traits. Both groups of methods are discussed in turn in this section.

2.4.1 *Mode of inheritance based methods*

A mode of inheritance for the trait (i.e. the conditional probability distribution [phenotype | genotype]) must be defined (either known or assumed), and the likelihood of the observed data, $\mathbf{X}$ (the set of phenotypes and marker genotypes within a set of pedigrees or relative-pairs) is considered under this mode of inheritance.

Two-point analysis

For two-point analysis, a trait locus is assumed to exist at a recombination distance θ from a marker locus where genotypes are known. The aim is to estimate θ. The likelihood of the observed data, conditional on the marker genotypes, mode of inheritance and θ is maximised over $\left\{\theta : 0 \leq \theta \leq \frac{1}{2}\right\}$, where θ is the recombination fraction between the marker and disease loci. The maximum likelihood estimate of θ, $\hat{\theta}$, can be used to test the null hypothesis of no linkage ($\theta = \frac{1}{2}$) using a likelihood ratio test. The test statistic used is the lod (log odds) function,

$$Z(\hat{\theta}) = \log_{10}\left[\frac{L(\mathbf{X}|\theta = \hat{\theta})}{L(\mathbf{X}|\theta = \frac{1}{2})}\right]$$

which may be interpreted as a comparison between the likelihood of observing the data under $\theta = \hat{\theta}$ versus the hypothesis of no linkage ($\theta = \frac{1}{2}$).

Multilocus and multipoint analysis

Traditionally, linkage analysis proceeded as a series of pairwise two-point tests between a trait locus and each of a number of marker loci. Advances in molecular biology have led to increasingly dense maps of markers across the human genome and multilocus and multipoint analysis is now common.

For multilocus methods, where the relative positions of several marker loci are known, all marker genotypes can be used simultaneously to find a lod score at each locus (Lathrop et al., 1984; Lathrop and Lalouel, 1984). This gives increased power over two-point methods (Lathrop et al., 1985). Multipoint methods go a step further by calculating the likelihood of the trait locus being at a location x, for a set of x spanning the marked region (i.e. including points between marker loci) and for an x unlinked to any marker loci in the region, often denoted as the point $x = \infty$. The location score at x is the log-likelihood ratio

$$2\ln\left[\frac{L(\mathbf{X}|x)}{L(\mathbf{X}|x = \infty)}\right]$$

and can be divided by $2\ln(10)$ to give a multipoint ‘lod’ score at x.

2.4.2 Model-free analysis of sibling pairs

Methods based on allele-sharing are often preferred for complex traits because no mode of inheritance need be specified. They are intuitively simple, based on a comparison of observed allele sharing at marker loci among affected individuals with that expected under the hypothesis of no linkage. Significant deviation is taken as indicative of linkage between the marker locus and a disease-influencing gene. They are often referred to as ‘model-free’ or ‘non-parametric’ methods, to distinguish them from the ‘lod-score’ or ‘model-based’ methods described above. However, these titles have been controversial, as discussed later in section 2.4.5.

Identity by descent and identity by state

Model-free methods fall into two groups: identity by state and identity by descent. Two alleles are *identical by state* (IBS) if they share the same DNA sequence and are *identical by descent* (IBD) if they also have a common ancestral source (i.e. descend from the same chromosome of the same ancestor). Two loci are increasingly likely to share a common IBD status as the degree of linkage between them increases and this allows the use of IBD-based statistics in detection of linkage. Any alleles that are IBD must be IBS (though the reverse does not hold) and IBS-based statistics have also been used in linkage detection.

Statistical detection of linkage based on IBS information is not as powerful as that based on IBD. IBS methods can suffer in comparison because they use less of the available information. On the other hand, IBD methods suffer when IBD status cannot be uniquely determined and individuals have to be left out of the analysis. Not only does this reduce power, but it can also lead to bias against linkage, since it is easier to identify siblings who share no alleles IBD than those who share one or two alleles IBD, and so these siblings will be overrepresented in the analysed subset. Methods to deal with this include typing more polymorphic closely linked markers and averaging any statistic used to detect linkage over all possible IBD states, weighted by the posterior probability of each state given the observed data. As marker polymorphism increases, the IBS distribution approaches IBD.

		Trait A		
		like	unlike	total
Trait B	like	n_0	n_1	$n_0 + n_1$
	unlike	n_2	n_3	$n_2 + n_3$
	total	$n_0 + n_2$	$n_1 + n_3$	n

Table 2.1: Example table for Penrose's sib pair method of linkage analysis

IBS methods

One of the first allele sharing methods of linkage analysis was proposed by Penrose (1935). We can tabulate sib pairs in a 2x2 table, according to whether they are alike or not for two traits, as shown in table 2.1.

Assuming no linkage between the genes controlling traits A and B, sibling pairs who are concordant for A should not be distributed in any way differently for trait B than those who are discordant for trait A. (For disease mapping, we would take trait A to be the disease phenotype and trait B to be the marker genotype). Significant deviation of the observed counts of sib pairs in the table cells from that expected under Mendelian segregation is indicative of linkage between the genes controlling the two traits. This can be tested using the usual χ^2_1 statistic.

IBD methods

Various statistics based on counting the alleles shared IBD between sibling pairs have been proposed. The *affected sib-pair method* (Cudworth and Woodrow, 1975) is similar to Penrose's, but with some improvements: marker sharing is defined by IBD status (not IBS) and attention is restricted to sib pairs in which both members are affected. Assuming no linkage between trait and marker loci, affected sib pairs would be expected to share 0, 1 or 2 marker alleles IBD in the proportions $\frac{1}{4} : \frac{1}{2} : \frac{1}{4}$. If n sibling pairs are recruited, and n_0, n_1 and $n_2 (= n - n_0 - n_1)$ are observed to share 0, 1 and 2 alleles IBD at a locus, then deviation from the expected numbers $e_0 = \frac{n}{4}, e_1 = \frac{n}{2}$ and $e_2 = \frac{n}{4}$ is taken to be indicative of linkage and can be tested using Pearson's χ^2 statistic:

$$S = \sum_{i=0}^{2} \frac{(n_i/n - e_i)^2}{e_i} \sim \chi^2_2.$$

One degree of freedom (1 df) tests are generally more powerful than 2df tests, and Suarez et al. (1978) suggested comparing just the proportion of siblings sharing two alleles IBD with the expected proportion, 1/4, using the statistic

$$\frac{(n_2/n - e_2)}{e_2}\left(\frac{n}{3}\right)^{\frac{1}{2}} \sim t_{n-1}$$

and a one-sided test, since we are only interested in the observed proportion exceeding the expected. This statistic is most powerful for recessive traits (Whittemore and Tu, 1998) because of the focus on excess sharing of *both* alleles at a locus.

The 1df family of test statistics of the form $T = \nu n_1 + n_2$ was explored by Knapp (1995). Setting $\nu = \frac{1}{2}$ gives the 'mean test' statistic (T is then half the mean number of alleles shared IBD in the sample), which is asymptotically normal with mean $\frac{n}{2}$ and variance $\frac{n}{8}$ which leads to the statistic

$$\frac{n_2 + n_1/2 - n/2}{\sqrt{n/8}} \sim t_{n-1}$$

for testing linkage. This was shown to be locally most powerful for all possible single-locus models and uniformly most powerful when $\frac{n_0 n_2}{n^2} = (\frac{n_1}{n})^2$.

All these tests, however, require that IBD status can be unambiguously determined at the locus under test. These tests have therefore often been restricted to families that are fully informative for IBD.

Multiplex sibships

The above tests are designed for sibling pairs, but often affected trios and above will be recruited and must be broken up into pairs. Suppose we have collected an affected sibling trio $(1, 2, 3)$. This can be split into pairs by

1. choosing one pair from each sibship (eg use pair $(1, 2)$ and discard sibling 3);
2. using all independent pairs (eg use pairs $(1, 2)$ and $(1, 3)$); or
3. using all possible pairs (ie use pairs $(1, 2)$, $(1, 3)$ and $(2, 3)$).

In the first option, there will be a loss of power due to discarding some data, and in both the first and second options the result may depend on the choice of pair. Therefore, the third option, using all possible pairs, is often preferred.

Using non-independent observations, though, can increase the false positive rate. To account for this, each observation can be weighted. One common scheme is to weight each pair from a sibship of size n by $2/n$ (Suarez and Eerdewegh, 1984), which means the weighted number of pairs is $n-1$, which is the number of independent pairs that could have been formed.

In fact, using all pairs does not inflate the false positive rate for counting methods (Blackwelder and Elston, 1985) though when using likelihood ratio methods the false positive rate may increase (Abel and Müller-Myhsok, 1998) or decrease (Meunier et al., 1997), depending on family structure and marker informativeness. Both studies found $2/n$ to be a conservative weighting scheme.

An alternative method of linkage analysis, which does not require splitting a multiplex sibship but, rather, deals naturally with sibships of all sizes was proposed by de Vries et al. (1976). This method tests for non-random segregation of haplotypes based on the distribution of the number of affected siblings receiving a particular parental haplotype (in favour of the other) from their parent. For the ith sibship of size n_i, let parent j ($j = 1, 2$) have genotype be $A|B$ and let n_A and n_B be the number of siblings who received haplotype A and B respectively. Then $D_{ij} = |n_A - n_B|$ is the observed difference between the number of affected siblings with one and those with the other parental haplotype. The test is based on comparison of D_{ij} with the expected difference, d_{ij}. Under random segregation, $n_A \sim B(n_i, 1/2)$ and so $d_{ij} = E(D_{ij})$ and $\sigma^2_{ij} = V(D_{ij})$ may be calculated. D_{ij} and d_{ij} are compared using the statistic

$$N = \frac{(|\sum D_{ij} - d_{ij}| - 0.5)^2}{\sum \sigma^2_{ij}}$$

which has a χ^2_1 distribution.

Note that this is a quite different test to the transmission/disequilibrium test (TDT) proposed by Spielman et al. (1993). In the latter, the authors test for non-random segregation of *particular* haplotypes using parent-child trios, ie testing for linkage in the presence of association. The method proposed by de Vries et al. (1976) tests for non-random segregation of HLA haplotypes in general among affected children of each parent, ie testing purely for linkage. It can be shown to be equivalent to the mean test for $n_i = 2$ (Abel et al., 1998a).

2.4.3 Likelihood based analysis of relative pairs

Further to the work described in section 2.3, Risch (1990b,c) proposed a method for linkage analysis known as MLS (Maximum Lod Score). The likelihood of observed genotypic data in affected sib pairs can be expressed as a function of the proportion sharing 0, 1 or 2 alleles IBD, $\mathbf{z} = (z_0, z_1, z_2)$. This can be compared using a likelihood ratio test to the likelihood of observed data under the hypothesis of no linkage (under which the sharing proportions are $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2) = (1/4, 1/2, 1/4)$):

$$\Lambda = \frac{\prod_{j=1}^{N}(\sum_{i=0}^{2} z_i w_{ij})}{\prod_{j=1}^{N}(\sum_{i=0}^{2} \alpha_i w_{ij})} \tag{2.2}$$

where w_{ij} is the probability of observing the markers of the jth pair, given that they share i alleles IBD. $\log_{10}\Lambda$ is a lod score which can be maximised over $\mathbf{z}$ to find posterior probabilities of IBD sharing at a locus. The maximised test statistic is

$$\text{MLS} = \max_{\mathbf{z}} \sum_{j=1}^{N} \log_{10}\left(\frac{\sum_{i=0}^{2} z_i w_{ij}}{\sum_{i=0}^{2} \alpha_i w_{ij}}\right).$$

subject to the constraint $\mathbf{z} \in [0,1]^3 : z_0 + z_1 + z_2 = 1$.

However, not all values values of $\mathbf{z} \in [0,1]^3 : z_0 + z_1 + z_2 = 1$ are compatible with genetic models. Risch (1990b) showed that, approximately, values of z could be expressed as functions of λ_R, with $z_0 = 1/4\lambda_S$, $z_1 = \lambda_O/2\lambda_2$ and $z_2 = \lambda_M/\lambda_S$. Expressing λ_R in terms of population additive and dominance variances gives the relations $\lambda_O \leq \lambda_S$ and $\lambda_M \geq \lambda_S$. Holmans (1993) used these relations to show that the possible sharing probabilities at any disease-locus must lie in the triangle in the (z_0, z_1) plane bounded by $z_0 = 0$, $z_1 = 1/2$ and $z_1 = 2z_0$ (see figure 2.3). Simulation showed that maximisation subject to these restrictions gave increased power to detect linkage over the unrestricted test.

The binomial method of de Vries et al. (1976) has also been formalised in a likelihood based statistic, known as the maximum likelihood binomial (MLB) method. The likelihood of observed data can be expressed as a prod-

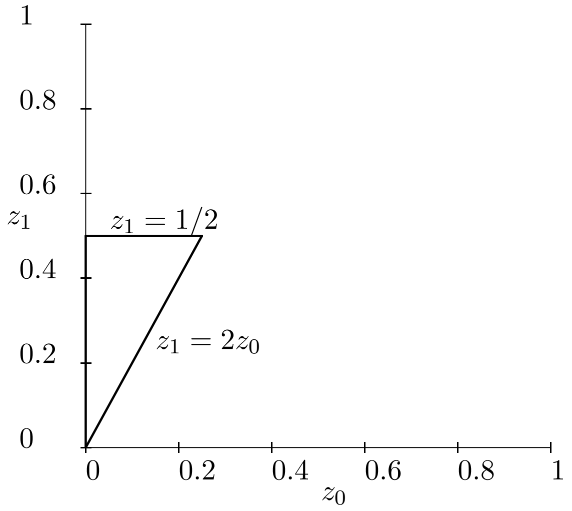


Figure 2.3: Holman's possible triangle for IBD sharing probabilities between sibling pairs

uct of binomial distributions (n_i, α) over parents (j) and families (i) where n_i is the number of affected offspring in family i and α is the probability affected sibs inherit the same allele from a given parent. Under the null, $\alpha = \frac{1}{2}$, and so the standard likelihood ratio test statistic

$$\lambda = 2 \ln \left(\frac{L(\alpha = \hat{\alpha})}{L(\alpha = \frac{1}{2})} \right)$$

(where $\hat{\alpha}$ is the maximum likelihood estimate of α) may be referred to a χ^2_1 distribution. Simulation has shown that this test has accurate type I error rates and is at least as powerful as the mean test under single locus models (Abel et al., 1998a) and that it is robust to missing parental data (Abel and Müller-Myhsok, 1998).

2.4.4 Model-free analysis of extended pedigrees

IBS methods

By removing the restriction of IBD, Weeks and Lange (1988) generalised the affected sib-pair method to pedigrees - the affected pedigree member (APM) method. This is based on comparison of the IBS sharing among all affected relative pairs with that expected under Mendelian laws and assuming no linkage. They define Z_{ij} to be a statistic measuring the marker similarity between individuals i and j (eg the proportion of alleles shared

IBS) in a pedigree v, and take Z_v to be a (possibly weighted) sum of Z_{ij} over all affecteds in the pedigree. Statistics from different pedigrees may be combined in

$$T = \frac{\sum_v w_v[Z_v - E(Z_v)]}{\sqrt{w_v^2 \operatorname{var}(Z_v))}}$$

where w_v is a pedigree-specific weight. T can be assumed to have a standard normal distribution (using the Central Limit Theorem) for a large number of pedigrees. This was extended to a multilocus test statistic by Weeks and Lange (1992) who defined

$$Z_{ij} = \sum_{m=1}^{N} Z_{ij}^m$$

where m ranges over N marker loci and proceeding as before.

However, this statistic does not make use of all available information - other pedigree members, even when available, are not used to resolve IBD status and it is only multilocus, not multipoint analysis. APM has been shown to perform less well than most other non-parametric methods in a simulation study (Davis and Weeks, 1997).

IBD scoring tests

Whittemore and Halpern (1994a) describe a class of tests in which a score, S, is assigned to each possible pattern of IBD marker allele sharing, and this score averaged over all patterns consistent with the observed marker data and pedigree shape, with high scores corresponding to high allele sharing among affected relatives. The advantage here is clear - where IBD sharing can be unambiguously determined, this is used directly in the analysis. Where it cannot, an expectation is taken and individuals still contribute to the analysis. The contribution to the overall statistic for each pedigree is the difference between the score given observed marker genotypes and the relationship between affected members and that expected given only the relationship. Different score functions give rise to different tests. Two suggested in the paper are S_{pairs} and S_{all} (terminology from Kruglyak et al., 1996).

S_{pairs} is a function of the number of alleles shared IBD by each relative pair, summed over all affected relative pairs. It is defined as

$$S = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} f_{ij}$$

where f_{ij} is one fourth the number of alleles shared IBD by affected individuals i and j (so $f_{ij} \in \{0, 1/4, 1/2\}$).

However, when considering larger sets of affected members, it may be more impressive to find that a group of affected individuals share the *same* allele IBD rather than that they each share *some* allele IBD with each other. S_{all} is defined as follows. Let s_{i1}, s_{i2} be labels for the paternal and maternal alleles of individual i. Label all alleles at a locus across a pedigree such that two alleles get the same label if and only if they are IBD. Let $\mathbf{u} = (u_1, ..., u_n)$ where u_i is either s_{i1} or s_{i2} for a pedigree containing n affected individuals (there are 2^n such $\mathbf{u}$ for each pedigree) and $h(\mathbf{u})$ be the number of nontrivial permutations of $\mathbf{u}$ that leave $\mathbf{u}$ unchanged. We expect $h(\mathbf{u})$ to be large when there is extensive IBD sharing among the n relatives' alleles. Define the score to be

$$S = 2^{-n} \sum_{\mathbf{u}} h(\mathbf{u}),$$

where $\mathbf{u}$ ranges over all 2^n possibilities.

Both scores are expected to be high when IBD sharing is high and low when IBD sharing is low. They are thus used to detect linkage by looking for regions of high IBD sharing between affected individuals. S_{all} is a powerful score when all affected individuals share the *same* allele IBD and is more sensitive to single allele sharing than S_{pairs}, while S_{pairs} is more sensitive to two allele sharing (as may occur in a recessive trait).

Using either score, the test statistic for pedigree i is then defined as

$$Z_i = \frac{S_i - E(S_i)}{[V(S_i)]^{1/2}} \tag{2.3}$$

where the expectation and variance are taken over the distribution of the score conditional only on the relationship between affected individuals. The scores can be combined across pedigrees by using a weighted sum

$$Z = \frac{\sum_i \gamma_i Z_i}{\sqrt{\sum_i \gamma_i^2}} \tag{2.4}$$

which has mean 0 and variance 1. Under the Central Limit Theorem, $Z \sim N(0,1)$ when a large number of relative pairs of the same type are observed.

When observed marker data do not fully determine IBD sharing, an expectation can be taken, with S_i and Z_i in equation (2.3) replaced by

$$\overline{S}_i = E(S_i|\text{data})$$

and

$$\overline{Z}_i = \frac{\overline{S}_i - E(S_i)}{[V(S_i)]^{1/2}}$$

respectively, and Z in equation (2.4) replaced by

$$\overline{Z} = \frac{\sum_i \gamma_i \overline{Z}_i}{\sqrt{\sum_i \gamma_i^2}}.$$

The expectation of $\overline{Z}$ is still 0, but the variance will, in general, be < 1. This test is implemented in the software genehunter (Kruglyak et al., 1996), but $\overline{Z}$ is still compared to a standard normal, although the inaccuracy in the variance is acknowledged. The authors call this the 'perfect data approximation' and note that it will result in conservative p values - although just how conservative these might be will depend on marker spacing and polymorphism and on which pedigree members are genotyped.

To deal with this, Kong and Cox (1997a) proposed a likelihood model. Each IBD configuration is given a score, as before, and the probability of each configuration w in pedigree i is written $P_i(w|\delta)$ with δ a free parameter such that $\delta = 0$ corresponds to H_0 and $\delta > 0$ corresponds to an alternative of excess sharing. They show there is a class of models for which the log-likelihood $l(\delta) = \ln[P(\text{data}|\delta)]$ can be written in terms of the $\overline{Z}_i$ from equation (2.3).

They propose two specific models for $P_i(w|\delta)$: the linear model and the exponential model (see Kong and Cox, 1997a, for detailed definitions). In fact, the linear model does not lead to a real probability distribution, and the exponential model is preferred. These statistics are implemented in a modified version of genehunter, genehunter-plus (Kong and Cox, 1997a). The advantage of this model is that p values are not conservative.

2.4.5 Comparison of model-based and model-free methods

This review has focused mainly on model-free statistics, because they will be used in this project. The relative merits of model-based or model-free methods has been the subject of much debate in the literature. The main points are as follows:

Effects of misspecified parameters

Since it is unlikely that the genetic model for a disease will be known exactly, it is important to understand how model based methods behave when the model is misspecified. This is particularly important when the analysis is applied to complex diseases.

The parameters involved in these methods include p, the frequency of the disease-related allele and f_{ij}, the penetrances for a person with genotype i/j. Misspecifying either of these does not appear to increase the false positive rate and has only a small effect on power, as long as penetrances are not incorrectly specified as complete (Xu et al., 1998). More important is misspecifying the dominance of the genetic model. This can lead to a large loss in power and biased estimates of θ (the recombination fraction between the marker and disease locus), particularly if a dominant model is misspecified as recessive (Clerget-Darpoux et al., 1986).

To allow for the application of model-based methods to complex disease (where the disease model is unknown) some authors argue for the analysis to be performed under different disease models. This introduces multiple testing which must be accounted for (MacLean et al., 1993). Another alternative is to maximise the lod score over single locus models. For example, Curtis and Sham (1995) maximise the lod score over the heterozygote penetrance (f_{ij}). However, it is not clear how these methods perform when the underlying disease model involves multiple loci.

Both model-based and non-parametric methods require specification of population allele frequencies when founders are untyped and both are sensitive to misspecification of these. When founders are not typed, their genotypes are estimated conditional on the population allele frequencies. If two affected relatives share a marker allele, the probability the sharing is IBD and thus the probability of linkage between this allele and the disease allele increases with its rarity. Therefore incorrect allele frequencies in either direction can influence the evidence for linkage if a founder's genotype is unknown, and may lead not only to reduced power but also increased false positive rates (Ott, 1992). If all founders are typed, however, allele frequencies do not play a part in the model. It is likely to be a bigger problem for multipoint methods and when markers are closer together, and errors in intermarker distances and the order of markers also have a great influence on the result, particularly when markers are very close together.

Power

Model-free methods are attractive when dealing with complex traits *because* there is no need to specify explicitly a mode of inheritance. This also means they often take a 'one test fits all' approach by testing for any deviation in the observed data from that expected under the hypothesis of no linkage. Model based methods test for a specific type of deviation, which means they have more power when the correct model is specified, and have been shown to be at least as powerful as model-free methods in this instance (Abreu et al., 1999).

Under the incorrect model, however, parametric methods can perform poorly in terms of power to detect linkage and result in biased estimates of the recombination fraction (Clerget-Darpoux et al., 1986), although some authors have argued that the method of averaging results over several possible models is still more powerful than nonparametric methods even after adjustment for multiple testing (Durner et al., 1999).

Statistical equivalence

It has been shown that many model-free tests are statistically equivalent to parametric tests under particular modes of inheritance. For example, Knapp et al. (1994) showed that the mean sib-pair test is statistically equivalent to parametric lod score analysis under a recessive model. Whittemore (1996) demonstrated that in the case of complete IBD information, the S_{all} statistic is the efficient score statistic corresponding to a parametric likelihood, which implies that a test based on S_{all} is equivalent to a test based on maximising a parametric likelihood. In effect, a so-called model-free analysis may be testing for linkage under a specific genetic model without acknowledging this fact.

It is important to be aware of this equivalence between 'model-based' and 'model-free' methods if using the latter. But several authors argue that this should not preclude their use (e.g. Kruglyak, 1997). Different model-free tests will be sensitive to different types of deviation from the null and so no single test will be optimal in all situations. The best option seems to be to pick one that is robust to differences in the way deviations arise, rather than one that is particularly sensitive to one type of deviation. Sengul et al. (2001) examined the performance of ten commonly used nonparametric statistics under 27 different genetic models for affected sibships of sizes two to five. They found S_{all} performed consistently well, being either the most powerful or nearly the most powerful under all models.

2.4.6 Exclusion mapping

Linkage analysis is generally concerned with evidence *for* linkage between a trait and marker gene. But the flip side is to examine evidence *against* linkage. Thus, regions of chromosome that are highly unlikely to contain any gene linked to the trait can be excluded from further analysis. In single point lod score analysis, for a given marker, all values of θ such that $Z(\theta) < -2$ (corresponding to a likelihood ratio of $1 : 100$ or less) are excluded. Similarly, in multipoint lod score analysis, regions are excluded on the basis that $Z < -2$.

Exclusion mapping can also be performed using Risch's MLS statistic, which is made dependent on a value for λ_S (eg $\lambda_S = \lambda^*$). Under no dominance variance,

$$\begin{aligned} z_0 &= \alpha_0/\lambda^* \\ z_1 &= \alpha_1 \\ z_2 &= \alpha_2/(2 - 1/\lambda^*) \end{aligned}$$

where z_i and α_i are the posterior and prior probabilities respectively of sharing i alleles IBD. The likelihood of the observed data at a locus under $\lambda_S = \lambda^*$ can be compared to the likelihood under the null hypothesis of no linkage using the likelihood ratio defined in equation (2.2). A lod score can be formed by summing $\log_{10} \Lambda$ across pedigrees at each position. Regions where the lod score is strongly negative (< -2) can be excluded from containing a locus responsible for a locus-specific λ_S of λ^* or greater. Note that such a region might still contain a locus responsible for a smaller effect.

2.4.7 Criteria for declaration of statistically significant linkage

In all the above methods, some statistic is calculated that summarises the statistical evidence for linkage. As this statistic increases, so it becomes increasingly unlikely that such a statistic was reached by chance alone and likely that linkage exists. As with all statistical tests, significant results are declared when the statistic being used to detect linkage exceeds some predefined boundary. The boundary is commonly chosen such that a spurious significant result would occur randomly 5% of the time (corresponding to 'a type I error rate of 5%', or 'significance at the 5% level'). But the choice of where to set the threshold to declare linkage is important.

Typically $Z(\hat{\theta}) > 3$ has been taken as indicative of significant linkage, and been shown to correspond to a type I error rate of 5% when a single marker is being tested. However, when conducting genome screens this must be raised to account for multiple testing. While increasing the threshold decreases the false positive rate, it also decreases the power to detect linkage. Bonferroni type corrections are not appropriate when dealing with dense maps of markers because the markers will be linked and so the tests will not be independent. Lander and Kruglyak (1995) argued that a threshold of MLS > 4 was appropriate. Other authors have argued for other thresholds, eg Suarez and Hampe (1994) suggest 3.2. Such thresholds are often calculated under particular assumptions (eg an infinitely dense marker map), and may not be appropriate for all studies. Sawcer et al. (1997) showed that the Lander and Kruglyak value was conservative in their study of linkage in sibling pairs affected by multiple sclerosis. Using simulation, they showed that a threshold of MLS > 3.2 corresponded to a 5% type I error rate in their study and recommended that simulation be more widely used to determine the significance of results.

In practice, when an MLS > 3 is observed, further closely spaced markers are typed in the region or more families are recruited to increase further the MLS and narrow the region of interest.

2.4.8 Efficient strategies for conducting a genome screen

When conducting a genome screen, one option is to type all individuals at closely spaced markers across the genome, but this is likely to be an inefficient strategy. The main aim of choosing an efficient strategy is to minimise genotyping without losing power. One common strategy is to use a two-stage screen, with a coarse grid of markers in the first stage. In the second, only those areas showing suggestive linkage are typed using a finer marker grid. Another is to initially screen only a subset of the available individuals and screen the remainder only in those areas showing suggestive linkage in the first stage. Holmans and Craddock (1997) showed that an efficient strategy will involve a combination of these two. A low threshold should be chosen for the first stage, to ensure true linkage is not missed.

Typing unaffected relatives increases information about IBD states, and thus power, particularly when markers are not highly polymorphic. However, Holmans and Clayton (1995) showed that it is more efficient to increase the number of affected sibling pairs in a dataset than their unaffected relatives, though the cost of additional genotyping must be weighed against the cost of recruitment. Typing unaffected relatives also increases the chance of detecting genotyping errors or mispaternities. But affected relative pairs who have no unaffected relatives available for genotyping should not be discarded from analysis.

2.5 Linkage analysis and infectious diseases

As mentioned in section 2.4, genome screen linkage analysis has been successful in the localisation of genes which affect human susceptibility to infectious diseases. Three diseases are considered here: schistosomiasis which was the first (Marquet et al., 1996, 1999), TB (Bellamy et al., 2000) because of its relationship to leprosy and leprosy itself (Siddiqui et al., 2001; Tosh et al., 2002; Mira et al., 2003).

2.5.1 Schistosoma mansoni

The first successful mapping of a locus controlling susceptibility to an infectious disease was a genome screen of pedigrees affected by schistosomiasis (Marquet et al., 1996, 1999). Schistomiasis is caused by eggs laid by *Schistosoma mansoni* (schistosome worms), which live for years in the mesenteric and portal veins of a human host. Infection occurs when humans wade or bathe in waters infested with schistosome larvae. A segregation analysis of infection intensity (measured by eggs per gram of tissue), adjusted for age, sex and water contact was performed on 20 Brazilian pedigrees. The results suggested a co-dominant major gene controlling infection, with the allele predisposing to high infection levels at a frequency of 0.2–0.25 in the population (Abel et al., 1991). A genome scan was then performed using parametric linkage analysis under this model with 142 Brazilian people from 11 informative families. The results showed strong evidence of linkage at two closely linked markers in the 5q31-q33 region ($Z = 4.74$, $Z = 4.52$) and confirmed the existence of a codominant major gene controlling infection intensity. The authors also noted that no other marker had a single-point lod score above 1.1.

2.5.2 Tuberculosis

Tuberculosis (TB) is caused by infection with *Mycobacterium tuberculosis*. Bellamy et al. (2000) conducted a two stage genome scan on affected sibling pairs from the Gambia and South Africa. In the first stage, 92 affected sib pairs and their parents were typed for 299 microsatellite markers and analysed using the MLS method. Seven regions showed potential evidence for linkage ($\text{MLS} > 1$). In the second stage, 22 markers from these regions were typed in a further set of 82 sib pairs and parents. Only two regions, on chromosome 15q ($\text{MLS} = 2$) and Xq ($\text{MLS}=1.77$) continued to show evidence suggestive of linkage. Even though a previous case-control study in the Gambia found 2 candidate genes, *NRAMP1* and *VDR*, to be associated with TB (Bellamy et al., 1998), markers near these showed only weakly positive lod scores. The authors argued that this could be because susceptibility to TB was under polygenic control, and that the regions found in the linkage analysis would have a stronger effect than the candidate genes *NRAMP1* and *VDR*. The evidence, though, is not clear, and a larger sample size, or case-control trials of candidate genes in the regions showing suggestive linkage would be necessary to confirm this hypothesis.

2.5.3 Leprosy

Siddiqui et al. (2001) also conducted a two stage genome screen, using sibling pairs from India affected with leprosy. In the first stage, 103 sibpairs and their parents were typed for 388 polymorphic markers. Weak suggestive linkage (MLS $>$ 1) was found in 28 regions, and 37 markers from these regions were typed in a further 142 sibpairs and their parents. Only one region on chromosome 10p showed a multipoint MLS $>$ 3, and a further eight markers were typed in this region on all families to increase marker density. The maximum multipoint lod score was found to be 4.09 ($p < 0.00002$) on chromosome 10p13, with an estimated locus-specific λ_S of 1.66. The authors argue that this shows that susceptibility to leprosy is under polygenic control (of HLA and a gene located near 10p13) and that it is possible to map genes controlling susceptibility to infectious diseases using obtainable sample sizes.

Regions with multipoint lod scores above 1 but below 3 were examined by typing additional microsatellite markers. The marker D20S115 was found to have a multipoint MLS of 1.29 across all families, but when families were divided according to region, the multipoint MLS was found to be 3.16 for 175 families from Tamil Nadu and 0.38 for 70 families from Andhra Pradesh (Tosh et al., 2002). The singlepoint MLS for the Tamil Nadu families for this marker was 3.48. The authors found no evidence for interaction between markers in 20p12 and 10p13, suggesting loci at these two locations act independently, although the power to detect minor interactions with only 175 families is low.

The finding of linkage among families from only one of the two regions studied was unlikely to be explained by differences in diagnoses, since the same criteria were used in each area. However, it is known that the Indian population consists of many groups with different origins and migration histories and the authors argue that this can explain the different findings in the two regions, claiming that risk alleles may be population-specific. Indeed, population-specific MHC risk alleles have been found in South India in studies of TB (Pitchappan et al., 1984) and psoriasis (Pitchappan et al., 1989). The authors do not say whether evidence for linkage differed between the two groups for the chromosome 10p13 locus.

A more recent study recruited 86 affected sibships (205 affected siblings altogether) from Vietnam and their parents (Mira et al., 2003). Clinical leprosy may manifest in different forms (this will be discussed in detail in section 2.6.1) and in the Indian study only two of the cases were of the multibacillary (MB) type; the remainder were paucibacillary (PB). In contrast, 56% of the Vietnamese cases were MB, reflecting the higher incidence of this type of leprosy in Vietnam. An entire genome screen was conducted using 388 highly polymorphic markers at 10cM spacing. Using the binomial likelihood method of (Abel et al., 1998a), eleven regions were found to show evidence for linkage, with multipoint LOD scores above 1. These regions were saturated with more markers, and statistically significant evidence (LOD=3.88) was found only in the 6q25–q27 region and the observed IBD sharing in this region corresponded to a locus-specific λ_S of 2.21. A further 208 simplex families were enrolled and TDT testing provided supporting evidence for this linkage, with Bonferroni-correct p values of 0.008 and 0.0018 for the markers D6S1035 and D6S305 respectively.

The authors also found non-significant evidence for linkage in, among others, the 6p21 region which contains the HLA complex (LOD=2.62) and the 20p12 region (LOD=1.13). However, no evidence was found for linkage to the 10p13 region (LOD=0.22).

The authors noted that their sample contained a far higher proportion of MB cases than the Indian study, and so divided their siblings into three groups (MB concordant, PB concordant and discordant) and conducted the same analysis for each group for the 10p13 and 6q25 regions. Evidence for linkage to 6q25 was found in all three groups and the IBD sharing proportions in this region were similar across all groups. However, for the 10p13 region, there was evidence for linkage only among the PB concordant siblings (LOD=1.74), despite the lack of evidence for linkage to 10p13 in the sample overall. The authors conclude that the 10p13 region is linked to susceptibility to PB leprosy, while the 6q25–q27 region is linked to susceptibility to leprosy per se.

2.6 Natural history and epidemiology of leprosy

Leprosy is a chronic infectious disease of man caused by *Mycobacterium leprae*. It probably originated in Asia, spread to Africa and Europe and appears to have been introduced repeatedly into the Americas in the 15th and 16th centuries. It has disappeared from wealthier populations in recent centuries (the last documented cases in the British Isles had onset in 1798).

Prevalence of registered cases has fallen with increased availability of short-course multi drug treatment after 1980 and the WHO reported there were 597,232 cases registered for treatment worldwide at the end of 2000, with 719,330 cases detected in the same year (WHO, 2002). Leprosy is now concentrated mainly in tropical belt countries, particularly India and Brazil. It is essentially a disease of the peripheral nerves but it also affects the skin and sometimes other tissues. Although rarely fatal, it is a significant cause of disability. The mode of transmission of *M. leprae* is still unknown - the currently prevailing view is that infection is spread human to human through nasal emissions, though skin to skin transmission has been supported by some researchers, and others have proposed environmental sources of *M. leprae* in soil or animal reservoirs.

M. leprae has an extremely slow doubling time - nearly two weeks. This may be partially responsible for the long incubation period, which averages between two and four years, although periods as short as three months and as long as 40 years have been recorded (Bryceson and Pfaltzgraff, 1990).

There is no sensitive and specific test for infection with the leprosy bacillus, and thus the patterns of infection are unknown. Several authors (e.g. Bryceson and Pfaltzgraff, 1990) argue that leprosy is likely to be similar to tuberculosis (which is caused by infection with another mycobacteria, *M. tuberculosis*), with disease occurring in only a small ($\sim$ 10%) proportion of infected individuals.

2.6.1 Clinical classification of leprosy

Among those who develop disease, clinical manifestations present over a wide spectrum, classified by histopathologists on the 5 point Ridley-Jopling scale (see table 2.2) from tuberculoid (TT) to lepromatous (LL). Borderline patients (BB) tend to be immunologically unstable and prone to shift toward either end of the spectrum. Lepromatous disease is characterised by an extremely high bacterial load (up to 10^9 bacilli per gram of tissue in the dermis) and low to non-existent cell-mediated immune response. Tuberculoid cases have very few bacteria and very strong cell-mediated response. This variation in bacterial load leads to another classification for leprosy as paucibacillary (PB) or multibacillary (MB), with PB corresponding to tuberculoid and MB to lepromatous disease.

Classification:	TT ......BT ......BB ......BL ......LL tuberculoid ....borderline ....lepromatous paucibacillary .... multibacillary
Bacterial load	low/undetectable ....high
Cellular immune response	high ....low/non-existent
Humoral immune response	low .... high

Table 2.2: Classification of leprosy on the Ridley-Jopling scale

2.6.2 Stages at which genetics may affect development of disease

According to current views, only a minority of those exposed to *M. leprae* develop disease and they present with clinical symptoms over a wide range. It is not yet clear why one person develops a different clinical type of disease to another, or even develops disease at all. One can therefore imagine two stages at which both environmental and host genetic and non-genetic factors may affect first whether infection occurs and secondly the development of clinical disease. This is shown schematically in figure 2.4.

2.6.3 Non-genetic factors associated with increased or decreased risk of leprosy

Development of clinical leprosy is affected by several factors. The main non-genetic factors are:

Sex

In Asia, leprosy is reported more frequently in males than in females, but this may be an artifact (women being less thoroughly examined). In many African populations, including Karonga District, it is reported more often in females. Type of disease also varies by sex and, in all populations, lepromatous disease is found more frequently in males than in females.

Immune history

Exposure to other mycobacteria, including environmental mycobacteria and, in particular, vaccination with Bacille Calmette-Guerin (BCG) reduces risk of disease in all populations, though the amount of protection conferred appears to vary between populations (Bryceson and Pfaltzgraff, 1990). BCG

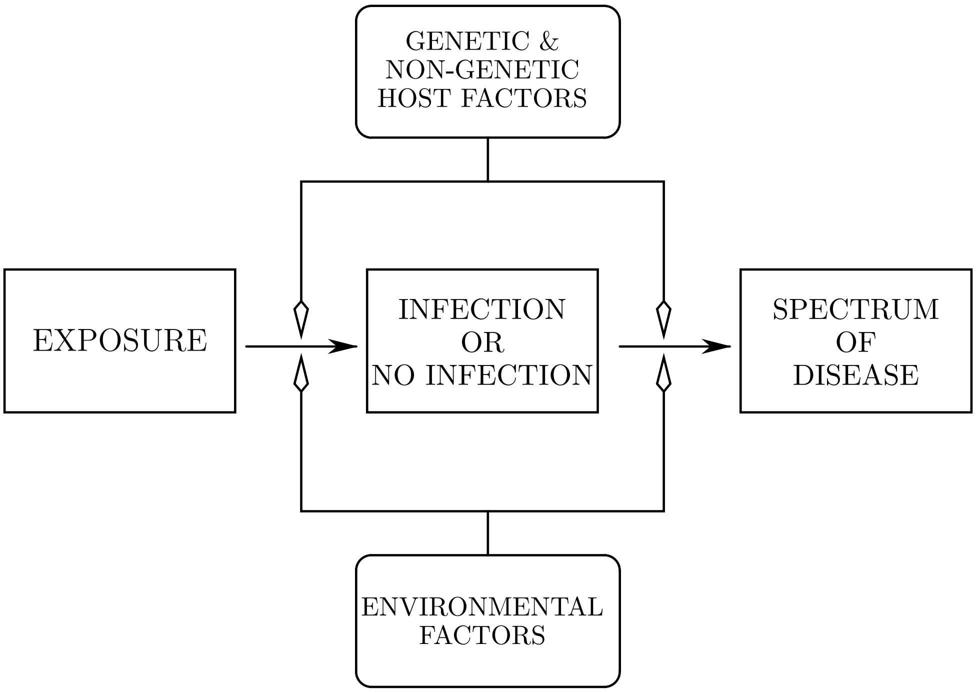


Figure 2.4: Schematic diagram of the stages when development of disease may be affected by host and environmental factors

vaccination has been shown to halve risk of disease in Karonga and it appears a second vaccination halves the risk of disease again (Pönnighaus et al., 1994).

Age

There is evidence in older literature (eg from Norway and the Philippines) that leprosy incidence was greatest in young adults in endemic populations (Fine, 1982). This pattern is no longer evident, however. Declining incidence and high uptake of BCG vaccination among children over the past several decades has shifted peak incidence to older age groups in most populations.

Exposure

Since leprosy is an infectious disease, exposure to the infectious agent is clearly necessary for disease. As mentioned above, however, the exact mode of transmission is not known. Studies have shown that sharing a household with an infected person substantially increases risk of disease, particularly if s/he suffers from MB leprosy (Fine et al., 1997).

2.7 Evidence for genetic influence on host susceptibility to leprosy

Several lines of evidence suggest that in addition to the above non-genetic risk factors, host genetics play a role in determining an individual's risk of developing clinical leprosy. Here, this evidence is considered, starting with the more general or cruder studies and leading to the more specific studies which implicate particular genomic regions or genes.

2.7.1 Susceptibility to atypical mycobacterial infections

Rare genetic mutations have been found which cause hypersusceptibility to weakly pathogenic mycobacterial infection. Such mutations confirm that genetic factors have an effect on susceptibility to mycobacterial disease in general, and indicate particular genes which are of importance for a healthy immune response to infection with mycobacteria. They have been found in genes coding for cytokines involved in immune response: IFNγ, which is part of the TH1 response leading to macrophage activation; and IL12, which plays a key role in influencing expression of IFNγ - see Marquet and Schurr (2001) for a full review.

2.7.2 Racial variation of susceptibility to leprosy

Although leprosy is found among people of all races, leprosy-type does appear to differ between races, with the proportion of lepromatous disease being lowest in Africans, higher in Asians, and highest amongst Caucasians. This difference has been shown to persist in migrant populations (Job, 1980), suggesting that it is related to host response rather than environment. However, some authors (e.g. Fine, 1988) suggest this may be due at least in part to variation in ascertainment of tuberculoid disease (characterised by hypopigmented patches which are most easily seen against a dark skin).

2.7.3 Twin studies

Monozygotic twins have identical genes, while dizygotic twins are expected to share only half their genes. On the assumption that both share a similar environment, higher concordance of a trait among monozygotic than dizygotic twin pairs is suggestive of genetic influence over the trait. Few twin studies of leprosy have been published and most suffer from poor study design. The most quoted was carried out in India by Chakravartti and Vogel (1973). Although clearly suffering from ascertainment bias (62 monozygotic and 40 dizygotic twin pairs were ascertained, although dizygotic twins are far more common), the results are at least consistent with a major genetic contribution to leprosy. Results are summarised in table 2.3.

Zygosity	Monozygotic	Dizygotic
pairs ascertained	62	40
both afflicted	37 (43.5%)	8 (20.0%)
type concordant	32 (86.5%)	6 (75.0%)

Table 2.3: Results of a twin study of leprosy in India by Chakravartti and Vogel (1973)

2.7.4 Segregation analyses

Segregation analysis aims to determine the transmission pattern of a trait within families and tests this pattern against predictions from specific genetic models - for example recessive or dominant. It should be noted that segregation analysis was originally designed to find single gene Mendelian traits, and has been successful in this area. However, the immune response to *M. leprae* is likely to be a complex trait, not under control of a single gene, and risk of disease is also affected by non-genetic factors. More recent segregation models claim to cope with these problems (see Jarvik, 1998) but several authors argue that segregation analyses are inappropriate for diseases such as leprosy (e.g. Cooke and Hill, 2001), often citing McGuffin and Huckle (1990).

McGuffin and Huckle used complex segregation analysis to ‘demonstrate the existence of’ a single major recessive gene for attending medical school - under a mixed model, a major recessive effect received more support than the multifactorial hypothesis. As McGuffin and Huckle acknowledge, while genetic factors may well play a part in the choice of medicine as a career, it is highly unlikely that it could be under control of a single gene. It is more likely this result is due to inadequacies of segregation analysis for the analysis of complex traits.

Despite this, some published segregation analyses have claimed to find evidence for particular genetic models for transmission of susceptibility to leprosy. Shields et al. (1987) describe an isolated population in Papua New Guinea within which the basic social unit was claimed to be the community and not the family. If this were true, the spread of an infectious disease in this population might be expected to be communal and not familial; but the reverse was found, with risk of disease associated with degree of relationship, suggesting genetic factors were influencing susceptibility. They also conducted a formal segregation analysis within one large kindred, but could not differentiate between any Mendelian genetic model, an unrestricted model and a purely environmental hypothesis.

Other published segregation studies have found evidence consistent with a major recessive gene affecting susceptibility to leprosy per se (Haile et al., 1985; Abel and Demenais, 1988; Feitosa et al., 1995; Shaw et al., 2001) and to tuberculoid disease (Haile et al., 1985; Abel et al., 1989). (Note that Haile et al. used data collected in an earlier study (Fine et al., 1979) which is discussed in the next section). Segregation analyses of lepromatous disease have been less conclusive (e.g. Serjeantson et al., 1979; Abel and Demenais, 1988) but tended to suffer from lower power because lepromatous tends to be rarer than tuberculoid disease. These analyses are summarised in table 2.4. None of these analyses were conducted in Africa and the only non-genetic factor accounted for was age.

2.7.5 *Specific regions of chromosome*

The search for evidence for the influence of particular regions of chromosome or specific genes over leprosy susceptibility has employed either linkage or association analysis, or a combination of both. Linkage analysis (discussed in detail in section 2.4) is concerned with identifying genomic regions containing a gene that affects a particular genetic trait. Typically, affected siblings and their parents, or multiply affected extended families are genotyped, and the genetic data are examined to test for regions that cosegregate with the trait more often than would be expected by chance.

Association studies use a case-control framework to test whether specific alleles of candidate genes are associated with a particular trait. Genes are

Author (year)	Population	Sample	Clinical Type	Conclusion[1]
Serjeantson et al. (1979)	Papua New Guinea	340 affecteds and 1st degree relatives	LL/LT	inconclusive; suggestive of a multi-factorial model
Haile et al. (1985)	India	72 pedigrees	all	major recessive gene, freq. ~ 0.02
			TT/BT	major recessive gene, freq. ~ 0.01
Shields et al. (1987)	Papua New Guinea	19 affecteds in single kindred (89 individuals)	all	NS
Abel and Demenais (1988)	Caribbean	27 multigenerational pedigrees	all	major recessive/ codominant gene
			TT/BT/BB	major recessive/ codominant gene
			LL/LT	NS
Feitosa et al. (1995)	Brazil	1568 families (10886 individuals)	all	recessive gene, freq. ~ 0.05
Shaw et al. (2001)	Brazil	76 families (1166 individuals)	all	two locus model, recessive major and modifier genes

[1] NS indicates no significant result

Table 2.4: Summary of segregation analyses for leprosy

typed in cases and (often matched) controls and the frequency of specific alleles compared between the two. If there is a significant difference observed, this indicates a particular allele is associated with increased or decreased risk of disease. A positive result may arise because the candidate gene is of functional importance in the disease process, because a gene in linkage disequilibrium with it is of functional importance or because the cases and controls were not adequately matched for genetic background. An important issue when conducting case-control studies is to ensure appropriate correction for the number of alleles tested. One very simple way this is done is to multiply any p value by the number of tests performed.

Such studies into susceptibility to leprosy are reviewed below, grouped by specific chromosome regions.

Human Leukocyte Antigen

The human leukocyte antigen (HLA) system is an extremely polymorphic region located on the short arm of chromosome 6 and is known to play an important role in cellular immune response, being responsible for presentation of specific antigens to T cells. The HLA region is within the Major Histocompatibility Complex (MHC) region which is divided into three classes which differ in structure and function and the cell types on which they occur. Figure 2.5 shows the organisation of the human MHC region.

Roles of MHC genes Genes within the class I and class II MHC regions encode the HLA class I and class II antigens. Class I molecules are found on the cell membrane of nearly all nucleated cells, while class II molecules are mainly present on immune system cells: macrophages and other antigen-presenting cells, B lymphocytes and activated T lymphocytes. The number of molecules of an MHC antigen may be altered by the presence of cytokines,

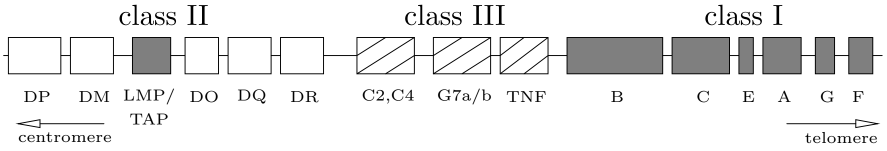


Figure 2.5: Organisation of the human Major Histocompatability Complex (MHC) region on chromosome 6

for example, the cytokine IFN-γ upregulates MHC class II expression. Genes encoding cytokines (among others) are located in the MHC class III region, which is located within one megabase of the MHC class II region.

The main function of class I and class II molecules in the normal immune response is to present antigen to T cells, which are unable to recognise antigens directly and can only participate in responses following MHC-associated presentation of antigen. This is known as *MHC restriction* and is one of the most important fundamental characteristics of the immune system.

Nomenclature of HLA class I and II alleles and antigens HLA typing was originally done using cell surface antigens, but is now often done using DNA based typing. As this change has happened, ability to discriminate between different alleles has increased and so the naming of HLA alleles and antigens has changed. Review of the HLA and leprosy association literature requires some understanding of the nomenclature of antigens and alleles within this highly polymorphic region. It is worth, here, giving a brief overview of the nomenclature used and to note that many antigen and allele names have changed over the years since the studies reviewed here began - see Marsh et al. (2001) for a complete report of current nomenclature.

The changing nomenclature reflects the history of the discovery of the HLA region and the discovery of new genes and antigens. The studies discussed in this section also reflect this history - earlier studies test only class I antigens (A, B and C), while class II antigens began to be studied as tests for them became field-robust. The HLA class III region has begun to be studied only in the past few years.

Each region (eg HLA-DR, a member of class II) contains several loci (eg HLA-DRB1, which codes for the β1 domain). Serological typing using alloantisera can detect class I and class II polymorphic antigens (eg HLA-DR2), but more recent cellular typing has detected new polymorphisms, including several *splits* of serological specificities (eg HLA-DR15 and DR16, which are splits of DR2). DNA typing has led to a further level of classification, and alleles are named according to the antigen they code for (eg HLA-DRB1*13 are the group of alleles at the DRB1 locus which encode the DR13 antigen). This information is presented in table 2.5.

Nomenclature	Indicates
HLA-DR	a region in the HLA system
HLA-DRB1	an HLA locus within the DR region, ie DRB1
HLA-DR13	an antigen encoded within the DR region
HLA-DR15(2)	a split of the HLA-DR2 antigen
HLA-DRB1*13	a group of alleles at the DRB1 locus which encode the DR13 antigen
HLA-DRB1*1301	a specific HLA allele within the HLA-DRB1 group (which codes for the DR13 antigen)

Table 2.5: Nomenclature of HLA alleles

Leprosy and HLA-I and II The binomial method developed by de Vries et al. (1976) has been employed in seven studies of genetic linkage to leprosy, summarised in table 2.6. These studies found strong evidence for non-random segregation of HLA haplotypes to affected siblings (they shared haplotypes more frequently than expected by chance). The evidence for non-random segregation was found to be stronger for tuberculoid siblings if parents were healthy, and stronger for lepromatous siblings if parents were also affected with lepromatous leprosy.

If a particular gene or haplotype conferred increased risk for leprosy per se, then its absence would be expected to confer some protection against leprosy. To examine this, many of these studies typed an older, healthy sibling and looked for evidence that they preferentially inherited a different HLA haplotype than their affected siblings. However, analysis showed no evidence of non-random segregation.

These observations led to a theory that HLA-linked genes do not affect susceptibility to *M. leprae* infection per se, but to the type of leprosy that develops once infection has taken hold, and that they act recessively towards tuberculoid disease and dominantly towards lepromatous disease. Although the results for the affected siblings are convincing, the apparent lack of disease in the healthy siblings and parents does not confirm that they were not susceptible. They may have had disease earlier and self cured, or might have later developed disease, so the dominant/recessive theory is less convincing.

Two other linkage analysis studies which applied ‘model-based methods’ (discussed in section 2.4.1) to HLA and leprosy (Haile et al., 1985; Abel et al., 1989) have been inconclusive. However, as will be discussed below,

Reference	Population	HLA class	Significant results[a] PB/PB	MB/MB	PB/MB
de Vries et al. (1976)	Surinam	I	↑	.	↓
Fine et al. (1979)	South India	I	↑	.	–
de Vries et al. (1980)	Central India	I II	↑	–	.
van Eden et al. (1980)	Central India	I II	↑	–	.
Bale et al. (1985)	India	I	↑	.	.
van Eden et al. (1985)	Venezeula	I II	↑	↑	.
Keyu et al. (1985)	China	I	–	↑	↓

[a] '.' denotes that the test was not performed, '–' a non-significant result, ↑ significantly more and ↓ significantly less sharing than would be expected under random segregation. Significance is at the 5% level.

Table 2.6: Summary results for analyses of HLA haplotype sharing in affected siblings

HLA-DR antigens are particularly implicated in studies of association to leprosy, and DR antigens were not typed in Haile et al. (1985), and typed only in very few study members in Abel et al. (1989). A more recent study used combined segregation and linkage analysis to test for linkage in the MHC region (Shaw et al., 2001), and found strong evidence of linkage to the HLA class II region ($p = 2 \times 10^{-6}$).

Population- and family-based association studies of HLA antigens and leprosy are summarised in table 2.7.5. Although most studies report a few significant results, only the class II loci HLA-DR and HLA-DQ have been repeatedly implicated.

Tumor necrosis factor alpha

The tumor necrosis factor alpha gene (TNF-α) is located within the MHC class III region. TNF-α is a pro-inflammatory cytokine produced mainly by macrophages and T lymphocytes which may affect host response to infection by stimulating effector mechanisms and promoting granuloma formation. Serum TNF-α levels are raised during reaction responses in leprosy patients (Santos et al., 2000) and a recent case-control study in India showed significantly higher ($p < 0.001$) TNF-α production in polar tuberculoid than polar lepromatous patients (Kaur et al., 2001).

A particular TNF-α promoter polymorphism at position -308 (a G/A

Table 2.7: Summary of leprosy and HLA antigen association studies

Reference[a]	Population	Sample	Antigens tested	Significant results[b]
$^{(P)}$Smith et al. (1975)	Phillipines	80 cases; 194 controls	A	none
$^{(P)}$Youngchaiyud et al. (1977)	Thailand	30 cases (20 MB; 16 PB)	A, B	↑ Bw40 **
$^{(P)}$Rea et al. (1976)	Mexico	92 cases; 315 controls	A, B	none
$^{(P)}$Ottenhoff et al. (1987)	Ethiopia	61 BT cases; 39 controls	A, B, C, DR, DQ	^{T}none
$^{(P)}$Takata et al. (1978)	Japan	60 cases (28 MB; 32 PB); 184 controls	A, B, Bw, Cw4	none
$^{(P)}$de Vries et al. (1980)	C India	15 PB, 16 MB cases; 36 controls	DR	^{T}↑ DRw2 ^{T}↓ DRw8 *
$^{(F)}$van Eden et al. (1980)	C India	15 nuclear families	A, B, C, DR	^{T}↑ DRw2 *
$^{(P)}$van Eden et al. (1981)	C India	78 sporadic TT/BT cases;129 controls	A, B, C, DR	none
$^{(P)}$Miyanaga et al. (1981)[c]	Japan	54 MB cases; 167 controls	A, B, C, DR, MT	^{T}↑ DR2 * ^{T}↑ MT1 *
$^{(P)}$Izumi et al. (1982)	Japan	369 cases (295 PB; 74 MB); 110 controls	A, B, C	^{L}↑ B7 ^{L}↓ Bw54
		112 cases (84 PB; 28 MB); 55 controls	DR, MT	^{L}↑ DR2 ** ^{T}↑ DR2 * ^{L}↓ DRw9 ** ^{L}↑ MT1 ** ^{L}↓ MT3 **
$^{(F)}$van Eden et al. (1985)	Venezuela	28 nuclear families	A, B, C, DR	^{L}↓ DR3 ^{L}↑ MT1

[a] Studies were either population-based (P) or family-based (F)

[b] ↑ and ↓ indicate significant positive and negative associations respectively. Where the result applies only to a particular clinical type of leprosy the prefix ^{L} denotes lepromatous and ^{T} tuberculoid subtypes. Significance is at the 5% level, with *, ** and *** indicating the 1%, 10^{-3} and 10^{-4} levels respectively.

[c] p values not corrected for multiple testing

Table 2.7: continued

Reference[a]	Population	Sample	Antigens tested	Significant results[b]
$^{(P)}$Schauf et al. (1985)	Thailand	67 cases (32 MB; 35 PB); 32 controls	DR, DQ	$^{T}\uparrow$ DR2 $^{T}\uparrow$ DQw1 *
$^{(P)}$Gorodezky et al. (1987)	Mexico	76 cases (30 TT; 46 LL); 100 controls	A, B, C, DR	$^{T}\uparrow$ DR3
$^{(P)}$Kim et al. (1987)[c]	Korea	157 cases (124 PB; 33 MB); 162 controls	A, B, C, DR, DQ	$\uparrow$ DR2 $\downarrow$ DR4 *** $\downarrow$ DRw53 *** $\uparrow$ DQw1 *** $\downarrow$ DQw3 ***
$^{(P)}$Rani et al. (1992)	N India	118 MB cases; 237 controls	A, B, DR, DQ	$^{L}\uparrow$ Bw60 ** $^{L}\uparrow$ DRw8 $\uparrow$ DQw1 ** $\uparrow$ DR2 ** $\uparrow$ DQw7 **
$^{(P)}$Rani et al. (1993)[d]	N India	94 cases (41 LL; 25 BL; 28 TT); 47 controls	DR2, DR3, DQ1	$\uparrow$ DR2 $\uparrow$ DQ1
$^{(F)}$Cervino and Curnow (1997) [e]	S India Egypt	72 pedigrees 15 families	A, B A, B, DR	$\downarrow$ B21 $^{T}\uparrow$ DR2 *** $\uparrow$ DR2 ***
$^{(P)}$Visentainer et al. (1997)	S Brazil	121 cases; 147 controls	A, B, Cw, DR, DQ	$\uparrow$ DR2 $^{L}\uparrow$ DR2
$^{(P)}$Wang et al. (1999)	China	69 cases (40 LL; 10 BL; 4 BT; 15 TT); 112 controls	B, DR2	$^{L}\downarrow$ B46 *
$^{(P)}$Joko et al. (2000)	Japan	93 cases (21 LL; 24 BL; 17 BB; 26 BT; 5 TT); 114 controls	DR, DQ	$\uparrow$ DR2 *** $\downarrow$ DR53 ** $\downarrow$ DQ4 **

[d] This study tested specific alleles that code for particular antigens, so p values for the antigens themselves are not available

[e] Reanalysis of Fine et al. (1979) and Dessoukey et al. (1996)

Reference	Population	Number (frequency, %) of TNF2 in MB cases		PB cases		controls	
Roy et al. (1997)	India	121	(7.0)	107	(2.8)	160	(2.8)
Santos et al. (2000)	Brazil	210	(9.3)	90	(14.4)	92	(16.3)

Table 2.8: Summary of association studies of TNF2 and leprosy

transition named TNF2) has been studied in relation to leprosy, although it is also controversial whether this particular polymorphism actually affects TNF-α production. Some studies have found it is associated with increased TNF-α levels (Kroeger et al., 1997; Wilson et al., 1997; Louis et al., 1998) but others have failed to find evidence for such association. (Brinkman et al., 1995; Stuber et al., 1995; Somoskovi et al., 1999; Kaijzel et al., 2001).

A case control study in a (mainly male) Indian population of 121 lepromatous patients, 107 tuberculoid patients and 160 controls found that the TNF2 allele was present in the lepromatous group at a significantly higher frequency than the controls ($p = 0.03$), with an ethnic group adjusted relative risk ratio of 2.5 (Roy et al., 1997). Allele frequencies in tuberculoid patients were found to be similar to those in the controls. This study also found that TNF2 was in linkage disequilibrium with HLA-DR3, but that the frequency of HLA-DR3 was similar in both the controls and the lepromatous group. HLA-DR2 was found to be associated with both lepromatous and tuberculoid leprosy, but was not in linkage disequilibrium with TNF2, and stratification of the analysis for HLA-DR2 implied that the association with TNF2 and with HLA-DR2 were independent, despite their proximity to each other.

In contrast, Santos et al. (2000, 2002) found the reverse: the frequency of the TNF2 allele was significantly *lower* among 300 leprosy cases ($p = 0.005$), particularly the subset of 210 MB cases ($p = 0.001$), than among controls in Brazil. The mean serum TNF-α levels in the Santos et al. study were found to be similar between patients carrying two copies of the TNF1 alleles and those carrying one or two copies of the TNF2 allele, and higher in both groups during reversal reactions than before reactions. Allele frequencies from the Santos et al. and Roy et al. (1997) studies are shown in table 2.8.

The results of the Santos et al. (2000) study appear to be confirmed by another Brazilian study (Shaw et al., 2001). They conducted a combined segregation and linkage analysis on 76 Brazilian families (1,166 people) and examined sporadic, single-gene and two locus models. They found the best fitting model to be a two locus model, with recessive major and modifier genes. Linkage analysis under this model showed strong evidence of linkage to the HLA class II ($p = 2 \times 10^{-6}$) and TNF-α ($p = 2 \times 10^{-5}$) regions. Extended TDT analysis found the TNF1 allele was linked and/or associated with leprosy per se ($p < 10^{-4}$) and further two-locus TDT analysis suggested protective (TNF1/LTA2) and susceptible (TNF2/LTA2) haplotypes.

*Natural Resistance Associated Macrophage Protein 1 (*NRAMP1*)*

Studies in mice have shown that growth of *Leishmania donovani*, *Salmonella typhumurium* and susceptibility to *Mycobacterium bovis* are under host genetic control. Early studies mapped the three implicated genes (known as *Lsh*, *Ity* and *Bcg* respectively) to the same location, and later studies showed that the three genes were in fact the same, and *Bcg/Lsh/Ity* was renamed *Nramp1* (natural resistance associated macrophage protein 1). *Nramp1* controls susceptibility to a variety of mycobacteria in mice, with a dominant resistant allele and recessive susceptible allele (Buu et al., 2000).

The human homologue *NRAMP1* is located on chromosome 2q35 and is a candidate gene in studies of human susceptibility to mycobacteria due to the known function of *Nramp1* in mice. It has been shown to be involved in human susceptibility to tuberculosis (see Newport and Blackwell, 1997, for a review) and there is some evidence that it also plays a role in human susceptibility to leprosy.

Abel et al. (1998b) used segregation and linkage analysis in 20 multiplex families (166 individuals) in South-East Asia to examine the role of *NRAMP1*. They found significant ($p = 0.02$) non-random segregation of *NRAMP1* haplotypes using the method developed by de Vries et al. (1976), but found a maximum lod score for a recessive model of inheritance of only 1.28 (for $\theta \sim 25\%$). Monte Carlo simulations provided p-values for this lod score of 0.017 and 0.006 for the entire sample and a Vietnamese subset (16 families) respectively. They concluded that it is likely that leprosy is under the control of more than one locus, and that a gene close to *NRAMP1* is implicated.

A recent case-control study typed 273 leprosy patients and 201 controls from Mali for polymorphisms previously associated with susceptibility to tuberculosis (Meisner et al., 2001). They found no association with leprosy per se, but found a particular polymorphism (the *NRAMP1* 3'-untranslated region 4-bp insertion/deletion polymorphism) was associated with leprosy type - heterozygotes were significantly more frequent among MB than PB cases ($p = 0.007$). The authors conclude variation in or near the *NRAMP1* gene may affect the clinical presentation of leprosy, possibly by influencing cellular immune response.

The Mitsuda test measures immune response against intradermally injected lepromin (killed whole leprosy bacilli) and has been used to infer a prediliction for tuberculoid versus lepromatous disease (negative Mitsuda is associated with lepromatous disease). Alcaïs et al. (2000) used Mitsuda test results to infer susceptibility or resistance to lepromatous leprosy among 20 nuclear families with leprosy in Vietnam and tested for linkage between the Mitsuda result and *NRAMP1* markers. They observed significant ($p < 0.002$) evidence of linkage, which was independent of whether individuals were affected by leprosy or not.

These results, however, have not been consistently replicated. Hatagima et al. (2001) found no evidence of linkage between the Mitsuda reaction and *NRAMP1* in a sample of 30 sibling pairs from São Paulo, Brazil; Roger et al. (1997) found no significant evidence for linkage of leprosy with *NRAMP1* in French Polynesia and Roy et al. (1999) (see below) found no significant evidence for association of leprosy with *NRAMP1*. Shaw et al. (1993) found consistent evidence *against* a leprosy susceptibility gene in linkage with three markers in the 2q33–q37 region. They analysed 17 two and three generation multicase families from Pakistan and Brazil using two point analysis of three restriction fragment length polymorphism (RFLP) markers with leprosy per se, tuberculoid disease (with lepromatous diseased individuals coded as missing on unaffected) and immune response under recessive and dominant models. Every lod score was under 1, and many, particularly for small recombination fractions were under -2, which is generally taken as strong evidence against linkage (see section 2.4.6).

Vitamin D receptor

The vitamin D receptor (VDR) gene is located at 12q12-q14. VDR is involved in regulating calcium metabolism, and is also an immunomodulator involved in suppression of inflammation. A population-based association study on 166 controls, 107 TT and 124 LL leprosy cases recruited in India by Roy et al. (1999) (mostly the same individuals as in Roy et al. (1997)) considered polymorphisms in this gene and at 4 sites in *NRAMP1*. They found a significant difference ($p = 0.005$) in VDR genotypes between lepromatous and tuberculoid cases but none in *NRAMP1* genotypes (no significant difference between either group and the controls).

Laminin-α2

Wibawa et al. (2002) examined 53 leprosy patients and 58 healthy contacts from Indonesia and genotyped them for a missense mutation (T7809C) in of the laminin-α2 gene. Laminin-α2 has been shown to be a specific mediator for *M leprae* to bind to the Schwann cell surface. Patients were divided into a lepromatous group (LL, BL and BB clinical types; 27 patients) and a tuberculoid group (TT and BT types; 26 patients). They found the mutation was in Hardy-Weinberg equilibrium in the healthy contacts and lepromatous group, but not the tuberculoid group. Further, while there was no significant difference between the frequency of the mutation in the contacts or the patient group as a whole, there was a significant difference between lepromatous and tuberculoid patients ($p = 0.025$) with the mutation appearing at much higher frequency in the tuberculoid than the lepromatous group (73% vs$\dot{3}$0%).

Genome screens

Two genome screens for susceptibility to leprosy have been conducted in South India (Siddiqui et al., 2001) and Vietnam Mira et al. (2003); these have already been discussed in section 2.5.3. Functional genes in regions indicated by these screens have not yet been identified.

2.8 Summary and discussion

2.8.1 Genetic susceptibility to leprosy

The human response to infection with *M. leprae* is heterogeneous. This response is affected by environmental factors, but there is evidence that host genetics may also play a role in controlling this variation. There may be two separate genetic influences: one influencing development of clinical leprosy per se; the other affecting the type of leprosy that develops once disease has taken hold.

Twin studies into the heritability of leprosy have generally been poorly designed and although the results of three (out of four) segregation studies of susceptibility to leprosy per se point to a major recessive or recessive/codominant gene, none of these took account of common exposure within families.

The MHC region (class II in particular) is the most studied region in relation to leprosy, but reviewing studies is complicated by successive changes in nomenclature, as described in section 2.7.5. The DR2 antigen appears positively association with both tuberculoid and lepromatous disease, but evidence for other antigens is less consistent. The two studies which tested for such a DR2 association, but did not find one were in Venezuala and Mexico (there has been no published evidence of DR2 association in these countries).

There is also an apparent negative association between DR3 and lepromatous disease, implying DR3 may be protective against lepromatous disease. HLA-DQ antigens have been typed less often, but there is some evidence that HLA-DQ1 is positively associated with leprosy per se. A positive association between the DQ1 antigen and disease has also been observed in four out of five studies which tested for it - in Northern India, Korea and Thailand (but not Southern Brazil).

However, the Indian genome scan (Siddiqui et al., 2001) found no evidence of linkage to this region. It is not clear why this might be, and is particularly surprising given that the study was adequately powered and conducted in an Indian population (most positive studies of HLA and leprosy have been among Indian populations). There have also been significant, although contradictory, results from studies of the TNF-α gene in the MHC class III region (see table 2.8).

Despite the number of studies undertaken with HLA antigens, no antigens other than DR2, DR3 and DQ1 show consistent results. After multiple testing is accounted for, the other HLA antigens shown to be significantly associated with a type of leprosy or leprosy per se should probably be discounted as random false positives. A meta analysis might be a way to increase power and evaluate which positive results are false, but care must be taken in this area. If cases and controls are poorly matched for ethnic background, differences in allele frequencies unrelated to disease can occur simply because allele frequencies differ between ethnic groups. Further, if there is genetic heterogeneity between populations (as discussed below), it would not be appropriate to combine data from different populations.

Human *NRAMP1* was considered a good candidate gene given that its mouse homologue, *Nramp1* is known to control susceptibility to mycobacterial infections in mice. However, results of linkage and association studies in humans are not consistent. Out of six linkage analyses of the 2q35 region reviewed in this document, two found significant evidence for linkage (Abel et al., 1998b; Alcaïs et al., 2000, both in Vietnamese populations), three found no significant evidence (Levee et al., 1994; Hatagima et al., 2001; Siddiqui et al., 2001, in French Polynesia, Brazil and India respectively) and one found significant evidence *against* linkage (Shaw et al., 1993, in families from Pakistan and Brazil).

Recent work has indicated that a particular TNF-α polymorphism, TNF2, could be either directly associated with leprosy or could be in linkage disequilibrium in different populations with a polymorphism that affects TNF-α production and thus development of leprosy. It is not clear why the two case-control studies of this allele (Roy et al., 1997; Santos et al., 2000) find significant associations between TNF2 allele and PB leprosy, but in opposite directions. The studies were conducted in very different populations: the Brazilian population is racially very mixed, while the Indian population is seen as being relatively homogeneous in comparison. The frequencies of the TNF2 alleles also differ between the populations - TNF2 had an allele frequency of 2.8% among controls in India, but of 16.3% among controls in Brazil, nearly six times higher. It is possible that the Santos et al. result is due to poorly matched controls. Or it could be that while TNF-α levels in serum do have some effect on leprosy, the TNF2 allele does not itself affect TNF-α production but is in linkage disequilibrium with another polymorphism that does. Further work is needed towards understanding whether and how the TNF2 allele affects serum TNF-α and how TNF-α affects development of clinical leprosy.

Other reported associations, such as those with VDR (Roy et al., 1999) and laminin $\alpha 2$ (Wibawa et al., 2002), have not been replicated in any other published studies and functional polymorphisms have yet to be identified in the regions showing evidence of linkage in the South Indian genome scan

(Siddiqui et al., 2001; Tosh et al., 2002).

2.8.2 Reasons for lack of correlation between studies

Although the weight of evidence supports a hypothesis that HLA-DR2 is associated with leprosy, there is no evidence for linkage to chromosome 6q from the recent Indian genome screen. Results for other genes are not consistent between studies, and there are several possible reasons for this:

false positives When n independent statistical tests are performed, each with a significance level of α, one of them will be positive under the null with probability $1-(1-\alpha)^n$. For this reason, when multiple tests are performed, p values should be adjusted and significant findings should be replicated in other samples or populations. One possibility is that leprosy is not under genetic control at all, and all apparently significant results have been false positives.

genetic heterogeneity There could be different genes that affect susceptibility to disease. Not all alleles associated with increased risk of disease might be present in all populations, and so association or linkage studies in different populations with find different results.

gene environment interaction Genetic susceptibility might only be expressed in the presence of particular non-genetic factors. To take a simple example, suppose people are either genetically responsive or non-responsive to BCG vaccination. Those people who are non-responders would appear to be more susceptible to leprosy only in a population where most people were vaccinated. If gene-environment interactions act, and environment is not accounted for in analyses, results may again differ between populations who live in different environments.

pathogenic differences If *M. leprae* itself differs between geographical regions, this could also lead to apparent differences between genetic analyses of the human population in those regions.

While multiple testing problems might mean some of the reported associations between leprosy and particular alleles are false, it is unlikely that all significant results have been false positives. In particular, HLA-DR2 is supported by so many studies that it would be hard to argue against this being a true effect. Further, if we accept that HLA-DR2 does play a role in controlling susceptibility to leprosy, then other positive results from the MHC region may be due to linkage disequilibrium with HLA-DR (linkage disequilibrium is a prominent feature across the HLA region (Bugawan et al., 2000)).

The theory of genetic heterogeneity would explain the significant but opposite results for the TNF-α gene in Brazilian and Indian populations and has been proposed before to explain this difference (Santos et al., 2000; Shaw et al., 2001). These results could indicate that some unknown gene in the MHC region, in linkage disequilibrium with TNF-α (and possibly other HLA genes) is involved in the control of susceptibility. There might be genetic heterogeneity at this locus, or the association of high risk alleles at this locus might be in different directions in the Brazilian and Indian population. This might also explain other apparently population-specific results.

CHAPTER 3.

EPIDEMIOLOGICAL DATA - DESCRIPTION AND PRELIMINARY ANALYSIS

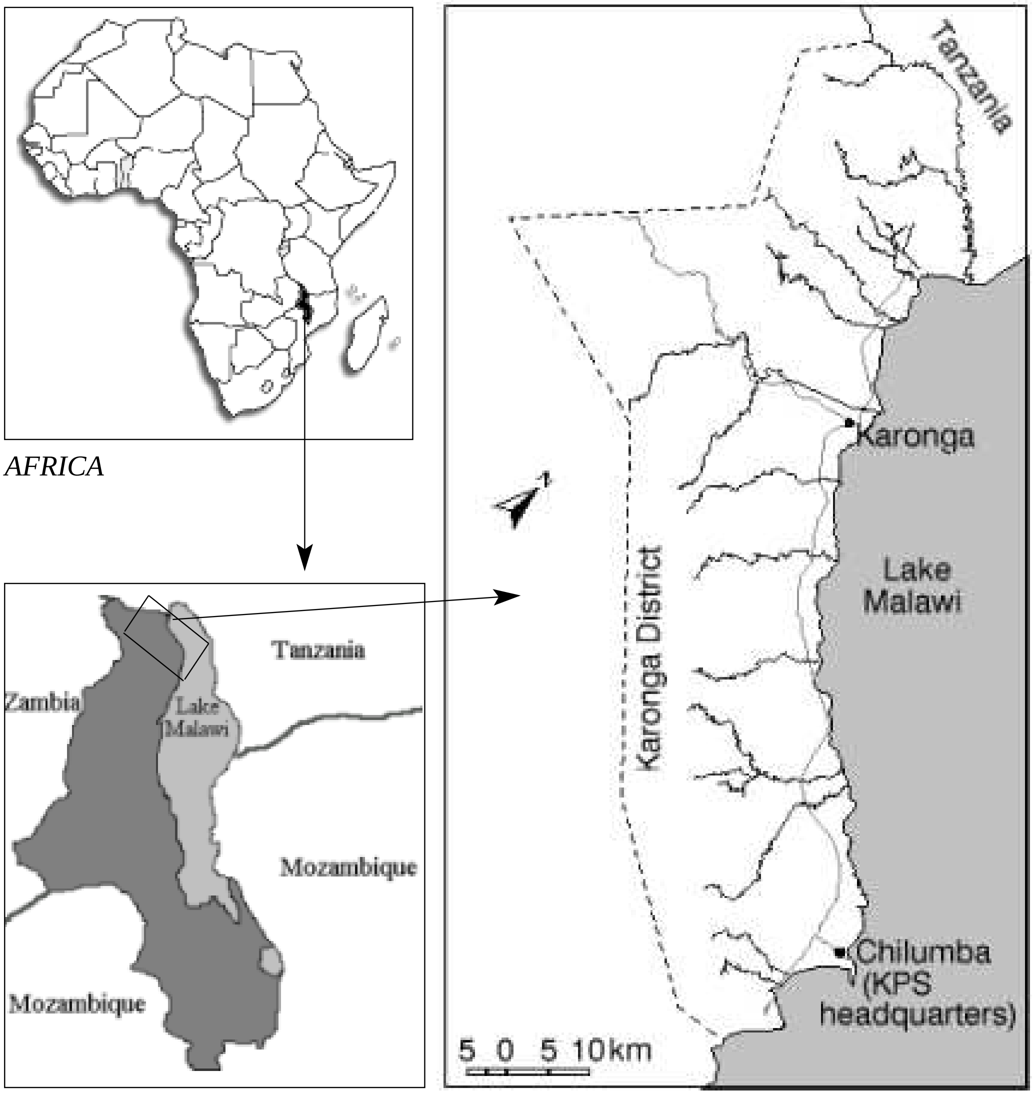
The data used in this project were collected by the Karonga Prevention Study (KPS). In this chapter, the history of the KPS and the relationship between the different surveys that have been conducted over the course of the study's history are described. Particular variables that will be used in this project are introduced and their collection and definition described. An exploratory logistic analysis of the cumulative incidence of leprosy among all individuals seen during the two complete population surveys is then presented. The effects of particular variables in the Karonga population are quantified and compared to their effects in other populations as described in section 2.6.3.

These data have been analysed extensively before within other KPS studies but the results of this chapter serve as an introduction to the data that will be used in this project and as a reference in later chapters to check the fit of more complicated models.

3.1 The Karonga Prevention Study

3.1.1 Background

The Karonga Prevention Study (KPS) is a major epidemiological study based in Northern Malawi which began in 1979, under the title of the LEPRA Evaluation Project (LEP). Karonga District lies along the shore of Lake Malawi as shown in figure 3.1. The initial focus of the KPS was leprosy, but this has broadened considerably and now includes tuberculosis (TB), HIV and helminths. Data have been collected on more than 250,000 living individuals to date, of whom 3,261 are or have been confirmed as leprosy cases, and 2,414 have had at least one episode of TB.



AFRICA

MALAWI

KARONGA

Figure 3.1: Map showing location of Karonga district, Malawi

3.1.2 Ascertainment and data collection

Two total population surveys were carried out in the 1980s (LEP1 in 1980–4 and LEP2 in 1986–9) during which all individuals in the area were visited in their homes by trained field staff from the local population. The two surveys covered virtually the same, but not identical geographical areas (eg a sparsely populated area in the western hills was covered in LEP1 but excluded from LEP2). Additional special surveys were carried out in small areas (within the larger study area) in 1984 and the early 1990s, to estimate leprosy incidence (as opposed to prevalence) prior to and during follow up of a vaccine trial. All surveys were carried out by field teams who conducted house to house visits and achieved very high coverage rates.

From 1989 to 1996, project staff were based at the hospital and peripheral clinics. All patients who attend any of these sites, for whatever reason, and who had not been seen for twelve months were screened for leprosy as in the surveys. Since late 1996, ascertainment has depended largely upon self-reporting, though several thousand individuals are examined each year by the KPS in the context of other studies.

Each individual was assigned a unique identifying number (six digits and one algebraically calculated checksum digit) and interviewers recorded individual data including year of birth, sex, parental identification and current household (each household is assigned a unique five digit household number). A household was defined as a group of people living together and acknowledging one person as its head.

Individuals were examined for leprosy by paramedical Leprosy Control Assistants (LCAs) and those who showed symptoms that may indicate leprosy were referred for a further review examination by a medical officer. Virtually all leprosy suspects were biopsied, and diagnostic certainty was assigned by an algorithm combining both clinical and histopathological findings (see figure in Pönnighaus et al. (1987) for a detailed description of this algorithm). Individuals were also asked about the time of onset for their symptoms, but it was recognised that the responses were often most unlikely (eg patients with old burnt out disease claiming very recent onset) and thus are not considered reliable (Pönnighaus et al., 1987).

3.1.3 Genetic studies

There have been two approaches to studying the genetics of disease in the KPS: an ongoing case-control study in leprosy and TB and a family study, of which this project is a part.

For the family study, each time a case of leprosy is identified, the database is screened to see if any (full) siblings of the case have also been diagnosed with leprosy. Once an affected sib-pair is identified, their pedigree is constructed up to fourth degree relatives and all affected people within the pedigree are identified. The following are selected for DNA collection:

- the affected siblings
- their parents; if one or both parents are not available, one or two unaffected siblings are also selected
- any other affecteds in the pedigree
- any people 'connecting' other affecteds to the sib pair or to each other, eg if an affected maternal affected cousin to the sib pair was identified, the parents of the sib pair, their grand parents and affected cousin's parents would also be selected.

Once someone has been selected for DNA collection, field staff in Karonga visit them and collect 7.5ml of blood in EDTA or a buccal swab sample from those people who prefer not to give blood. Not all people selected are available for DNA collection - they may have moved, died or refuse to give a sample. Some additional family members, not originally selected, have been bled because of their enthusiasm to participate. DNA is extracted in the project laboratory using Nucleon Kits, and is then sent to the Wellcome Trust Centre for Human Genetics in Oxford for analysis in the laboratory of Professor Adrian Hill.

As of September 2002, a total of 3,366 people had been bled (across both the family and case-control studies), 617 of them confirmed certain or probable leprosy cases. There are 270 cases and 447 controls in the leprosy case-control study and 529 cases and 1,087 controls in a parallel case-control study of TB. Out of 183 nuclear families identified in the leprosy family study as containing at least two affected siblings, 91 contain at least two affected siblings from whom DNA samples have been collected. Linkage analysis of these sibships is discussed in chapter 6.

Data are also available on multicase families with TB in this population and there is potential to apply methods described in this thesis to TB also.

3.2 Selection of data to meet assumptions

Epidemiological analyses used in this project to estimate the relative recurrence risk (see chapter 4) use methods which require present-state data, often termed point prevalence data, and assume that these data have been collected under complete ascertainment. The assumption of complete ascertainment holds for the periods of the two complete population surveys, but would not hold during the follow-up studies. Therefore, it was decided to restrict the data used in this part of the project to those people examined by the KPS before the end of LEP2. In this period (1979–1989), a total of 172,758 individuals were seen, 2,945 of whom were identified as leprosy cases. For the other parts of the project (based on linkage analysis), all available data have been used (ie data collected between 1979 and now).

3.3 Definition and description of covariates

Time

The incidence of leprosy has declined dramatically in Karonga district over recent decades. Because we require present-state data collected under full ascertainment, the date of the last examination before the end of LEP2 will be recorded as the time an individual was last seen. Figure 3.2 shows the number of people seen by the KPS by the year they were last examined. Those seen during only the first survey (LEP1) were considered likely to be different from those seen during both. Although efforts were made to trace all those seen in LEP1, some (5%) had died and others could not be found, typically because they had moved out of the district. Those seen during LEP1 but not LEP2 are more likely to have been born earlier, or be poorer or male (since men are far more likely to migrate to find work). However, while the mean year of birth is lower in the LEP1 only group (1959) than the LEP2 group (1965), reflecting the five year gap between the two surveys, the mean age in the two groups is the same - ~ 22.5. All other covariates,

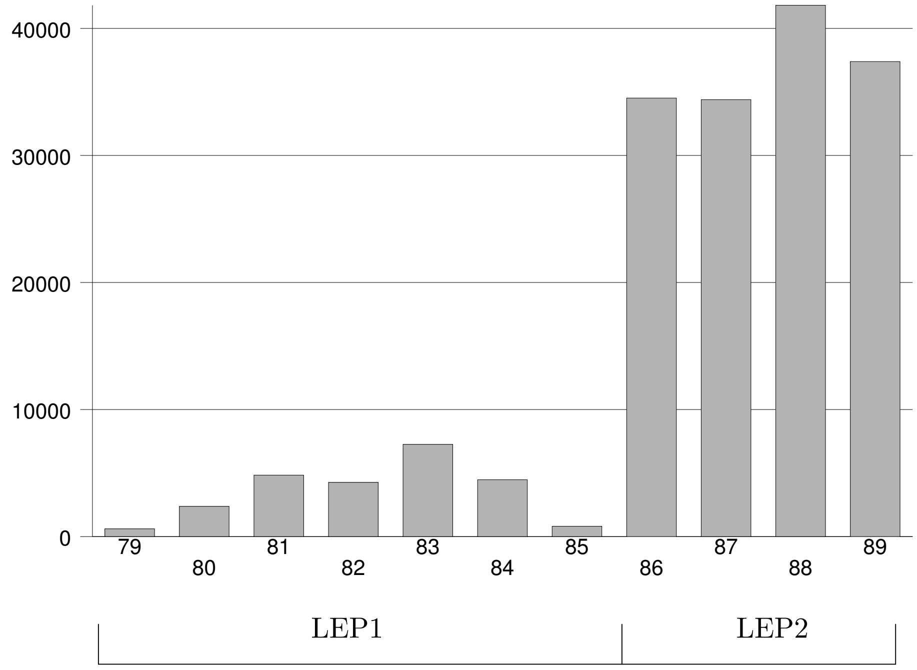


Figure 3.2: Number of individuals seen at least once by the KPS, by the last year they were examined

including the proportion ever affected by leprosy are similar, and therefore it was decided not to include any covariate to denote in which survey an individual was last seen.

Age/birth year

Of those seen, 15.8% did not give a precise year of birth. Instead, we have estimates on one of two scales: either by decade (1900–9, 1910–9, ... 1980–9); or as defined by a local events calendar; table 3.1 shows the calendar; used in LEP1. Estimated dates were used more often by women (22.6%) than by men (8.4%), and more often by older members of the population than younger.

Work in chapter 4 requires individual level data with ages in years. Excluding those who did not know their exact year of birth would decrease power and could introduce a bias, since these people are likely to be poorer and older and therefore at greater risk of leprosy. Three methods to deal with estimated birth years were considered:

Year	**Event**
1900	1900 or before
1914	Battle of Karonga
1934	Major crops damage by locust
1946	Passenger ship Viphya sank
1958	Dr Hastings Kamuzu Banda returned to Malawi
1964	Malawi gains independence

Table 3.1: Karonga local events calendar for the LEP1 population survey

Apportionment: For example, people who estimated they were born between 1910 and 1919, could be assumed to be distributed uniformly over this period, and hence contribute a count of 0.1 to each year $1910, 1911, \ldots, 1919$. This would be practical for analyses of aggregated data, but not for this project because, for some analyses, we require the data to be at individual level.

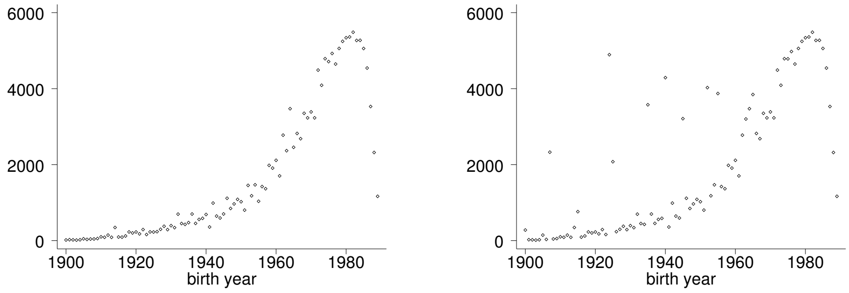
Midpoint assignment: Each person with an estimated year of birth could be assigned to the midpoint year of the estimated range. This distorts the distribution of people seen by the KPS by year of birth (see figures 3.3(a) and (b)).

Random assignment Each person with an estimated year of birth could be randomly assigned a single year within the range covered by the estimate. This could be done according to the empirical distribution of birth years among those who gave an exact year of birth across the range of years covered by the estimate. This method does not distort the distribution of people seen by the KPS by year of birth (see figure 3.3(c)).

It was decided to use random assignment as this preserved the age structure of the population. The age distribution of this population and the proportion of those who gave an estimated year of birth are shown in figure 3.4. In this project, age refers to an individual's age in years when they were last seen by the KPS before the end of LEP2.

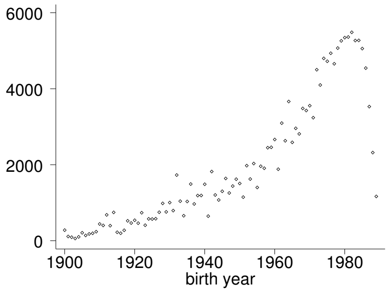
Sex

Sex was recorded for each individual seen, and the male:female ratio varies with age, with the proportion of females rising from 50.3% in the 0–9 age



(a) Population count by birth year: exact years of birth only

(b) Population count by birth year: midpoint assignment



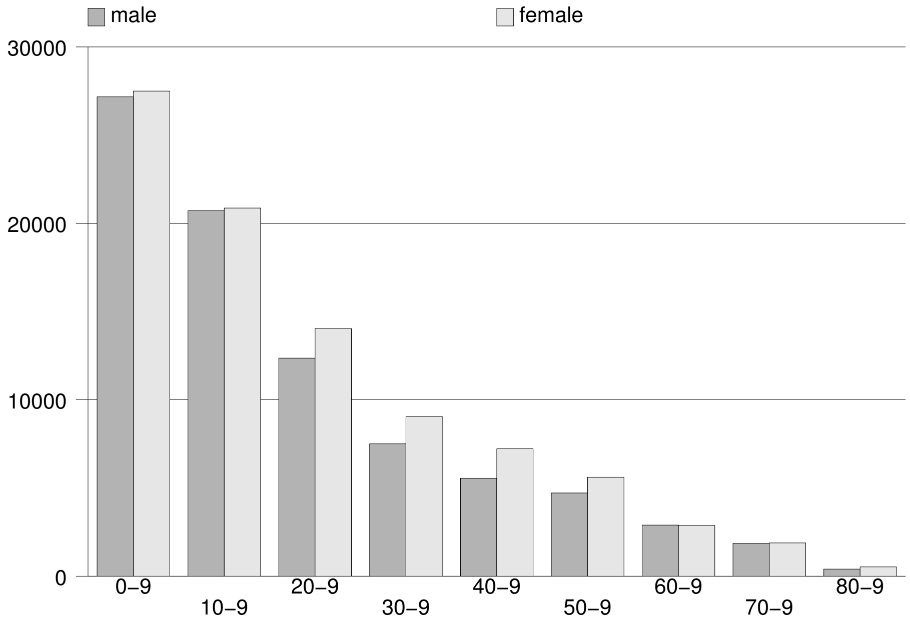
(c) Population count by birth year: random assignment

Figure 3.3: Distribution of individuals examined by the KPS before the end of LEP2: effect of midpoint and random assignment of birth years to those who gave estimated dates

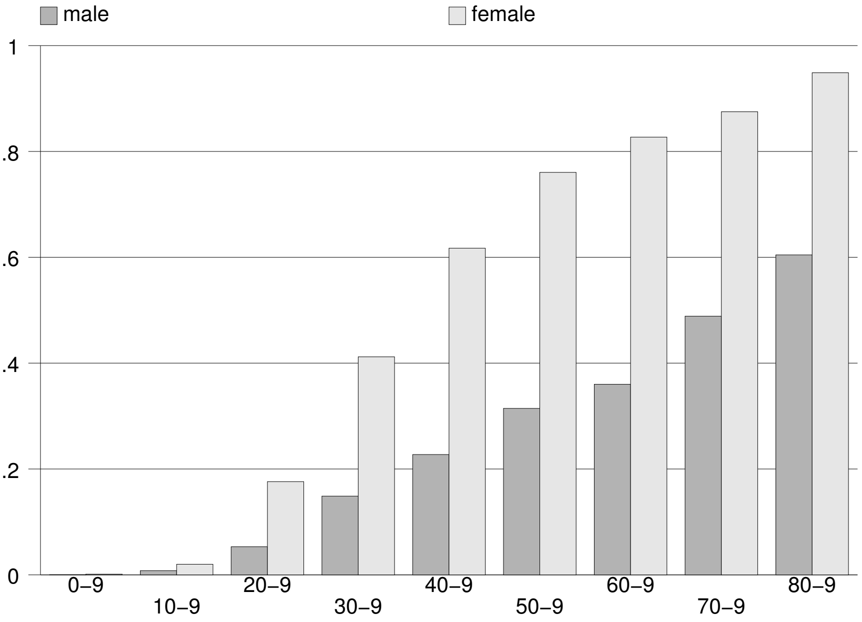
group to 56.6% in the 40–49 age group, before falling to 49.5% in the 60–69 age group and then rising again to 56.3% in the 80–89 age group. The first rise is probably due to men of working age working outside of Karonga district, while the second is probably due to the fact that women tend to live longer than men.

BCG scar

The older literature indicates that in populations without BCG vaccination, leprosy incidence peaks in 15–24 agegroup, but shifts to older groups as the incidence of leprosy falls. BCG vaccination is known to decrease the risk of leprosy (see section 2.6.3) and was introduced in Karonga in 1974 during



(a) Raw counts by agegroup and sex



(b) Proportion giving estimated years of birth by agegroup and sex

Figure 3.4: Age and sex distribution of individuals at their last examination before the end of LEP2

mass campaigns which initially targeted all inhabitants under 15 years old (particularly those in school). After three years, administration passed to infant vaccination services who made the vaccine available to infants in their first year of life. The KPS vaccinated many people of all ages in a trial between 1986 and 1989 and since 1990 the BCG coverage among infants is estimated to have been in the order of 75–80%.

Younger cohorts have thus had greater protection against disease, while, over the same period, exposure to the infectious agent has decreased (due to less contact with leprosy patients) and presumably other (socio-economic) factors associated with the disease have decreased across the population. This is likely to increase the shift of peak incidence to older agegroups which would be expected to change the pattern of cumulative incidence also.

For those individuals vaccinated by the KPS, vaccination history is known. For others, we use the presence of a BCG scar to infer vaccination. At each general examination, the presence of a BCG scar was noted as ‘yes’, ‘no’, or ‘doubtful’. Even when an individual has been vaccinated, a visible scar is not always present. Examination of BCG scars among 494 children in Karonga who were known to be vaccinated showed sensitivity of scar assessment decreases with time from vaccination, peaking at 95% at 7–12 months after vaccination and falling to 54% at 25+ months (Fine et al., 1989). Observations across multiple examinations were not always consistent. In this project, the variable `bcg` is coded as a categorical variable with three levels:

positive if a bcg scar had been consistently noted in general examinations before 1990;

negative if a bcg scar had been consistently noted as absent or if there were a series of absent and doubtful observations;

uncertain otherwise.

This means that all those coded bcg scar ‘positive’ will almost certainly have been vaccinated, but not all vaccinated people will be in this group. Therefore, while the effect of BCG vaccination is unlikely to be overestimated in this analysis, it may be underestimated.

The proportion of people with BCG scars varies by age and sex. Males are slightly more likely to have a scar (40% vs 37%), reflecting the higher

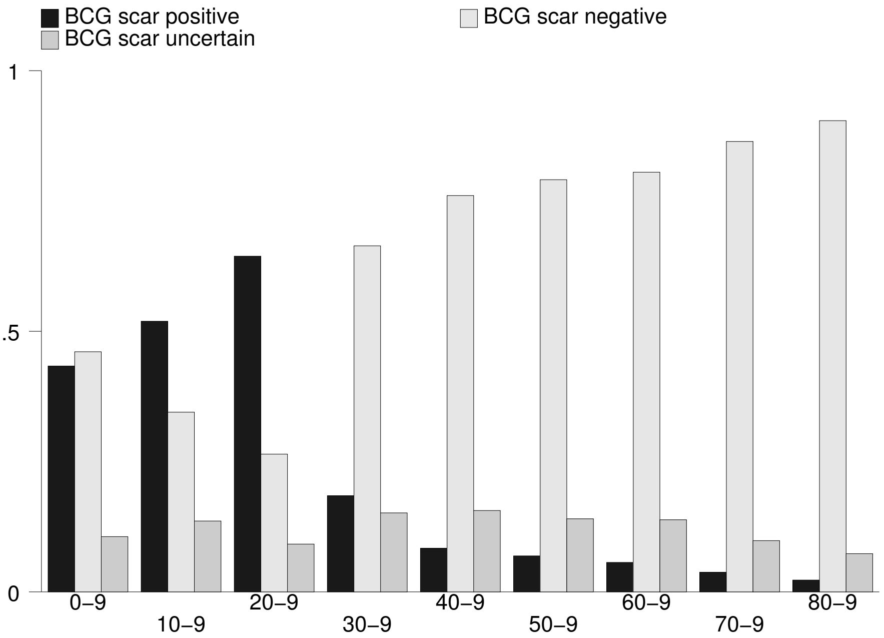


Figure 3.5: Proportion of people with BCG scar status 'positive', 'negative' or 'uncertain' at time of last examination before the end of LEP2 by agegroup

proportion of boys who attended school in the 1970s. The overall pattern by age is the same in both sexes, however, with rates rising towards the 20–29 agegroup, before falling - see figure 3.5. The proportion of those with uncertain scar status is broadly similar across all ages and both sexes: between 8% and 15%.

Household contact with leprosy cases

Exposure to *M. leprae* cannot be measured directly. In this project, household contact is used as a proxy for exposure. An individual is defined to be a household contact of a leprosy case if s/he shared the same household with a current (active) or past (cured) leprosy case during either of the LEP1 or LEP2 surveys. Such contact has been shown to increase the risk of leprosy substantially, particularly if that contact is with an MB case (Fine et al., 1997). Household contact in this project is coded by two variables: `pbcon` and `mbcon` are binary variables denoting household contact with a PB and MB case respectively. However, it is not possible to correctly identify all household contacts - people change households and migrate. This means some contacts are missed (Chirwa, 2001), which may again lead to the effect of household contact being underestimated.

3.4 Definition of main outcome variable of interest and further selection of data

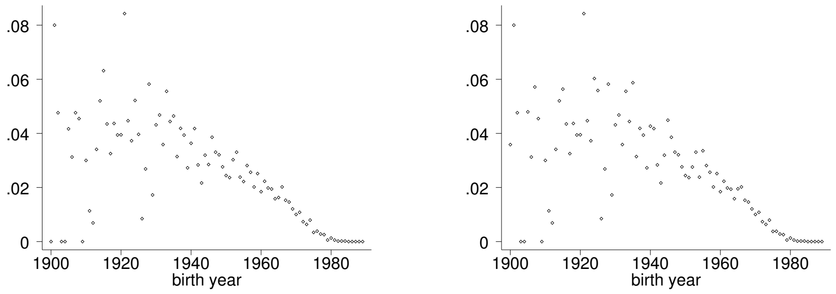
This project is about genetic susceptibility to leprosy. The main outcome variable of interest is *cumulative incidence of leprosy* - ie, did an individual have, or had they ever had, leprosy when they were seen by the KPS. All individuals who were classed as ‘certain’ or ‘probable’ cases by the algorithm described earlier will be considered cases in this project.

The effect of the random assignment of birth years to those who gave estimated dates is shown in figure 3.6. This shows that the method of random assignment of birth years to those who did not know their exact year of birth gave smoother variation in the proportion affected by birth year than if midpoint assignment had been chosen. It also had the effect of smoothing some of the variation among those born in the early 1900s wo knew their exact year of birth. This year on year variation was most likely due to the small denominator in this group (see figure 3.3(a)).

Incidence is low in older individuals. As described earlier, we expect incidence in younger age groups to have fallen as prevalence of leprosy has fallen and uptake of BCG vaccination has risen. We therefore expect cumulative incidence to rise with age, and that this rise will be considerably steeper among the agegroups who are more likely to become incidence cases. It is probable that this rise may level out after some cutoff age when incidence again falls, but it is unlikely that cumulative incidence would fall with age unless there were some age related problem with ascertainment.

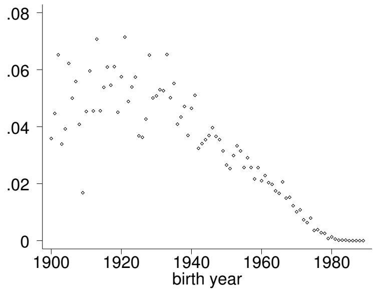
The proportion of people recorded by the KPS as ever affected by leprosy is shown by age in figure 3.7. There is some year on year variation due to random fluctuation, but grouping into 5 year age bands shows a clearer pattern. As expected, cumulative incidence is low below the age of 10 years, rises steadily to age 50 then remains about level till age 75 when it falls again. There is no evidence that older people were less likely to be affected by leprosy and this fall is likely to be due to artifact. Possible explanations include:

Mortality If leprosy were associated with increased risk of early death, older



(a) Proportion affected by birth year: exact years of birth only

(b) Proportion affected by birth year: midpoint assignment



(c) Proportion affected by birth year: random assignment

Figure 3.6: Distribution of cases among all individuals examined by the KPS before the end of LEP2: effect of midpoint and random assignment of birth years to those who gave estimated dates

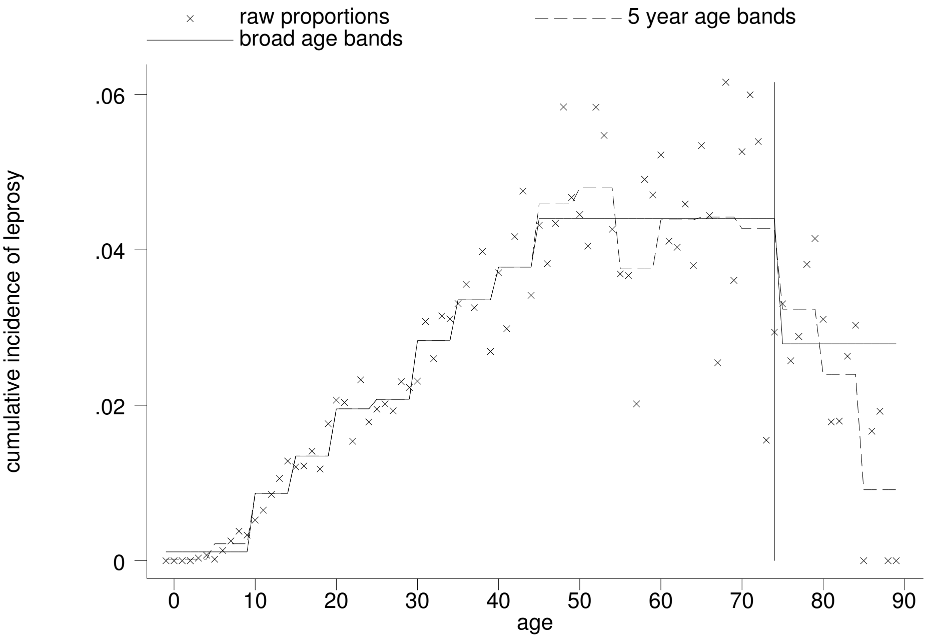


Figure 3.7: Cumulative incidence of leprosy among those examined before the end of LEP2 by age at last examination. The vertical line at age 75 shows the cutoff for inclusion in these analyses

cases might be missing from the population because they tended to die earlier. Leprosy is associated with lower socio-economic status which is itself associated with raised mortality. However, the relative risk of death for people affected with leprosy in this population has been estimated to be only 1.22 and not significantly different from 1 (Chirwa, 2001). Therefore mortality is unlikely to be the primary cause of this fall.

Ascertainment If disease in older individuals were missed, this could also lead to artificially low cumulative incidence. This may occur in older cohorts, because detection of skin lesions can be more difficult on older skin and neuromuscular deficit may be obscured, eg because of arthritis.

Self healing If someone had disease and then recovered with no skin lesions remaining before the KPS began, they would not be counted in the cumulative incidence rate. This is likely to be important since some individuals with milder forms of tuberculoid leprosy are known to selfheal - a study in Papua New Ginuea reported that during a period of 6 years when treatment was not available, 16% of affected individuals spontaneously healed (Shields et al., 1987).

Random variation The cumulative incidence rate is subject to random variation. Only a very small number of people (2,473) seen by the KPS belonged to the 75–89 agegroup, and, using the 0–9 agegroup as a baseline, the odds ratios for leprosy in the 45–74 and 75–89 agegroups are not significantly different at the 5% level (40.6, 95% CI 31.4–52.4 vs 25.3, 95% CI 17.9–35.7 respectively).

It is likely that the decrease in cumulative incidence in older age groups is attributable to a combination of all four factors. However, some of the work presented in chapter 4 assumes cumulative leprosy incidence is non-decreasing and 2,473 people aged 75 and over have been excluded from these analyses. This reduced the dataset further, to a total of 170,825 individuals, 2,876 of whom were known to have been leprosy cases.

3.5 Preliminary analysis

Figure 3.7 shows that the relationship between age and cumulative incidence of leprosy is non-linear. For subsequent analyses it was decided to treat age as a categorical variable, divided into nine broad bands: 0–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–74.

Although there appears to be a difference in cumulative leprosy incidence between all three BCG scar levels (positive, negative and uncertain) as shown in table 3.2, the difference between scar-negative and scar-uncertain groups is only borderline significant (just outside 5%) and becomes non-significant after adjustment for age. Therefore BCG scar status has been coded as a binary variable here, combining the BCG scar negative and uncertain groups.

The proportion of people giving birth estimates varies with age and sex (figure 3.4), with older people and women more likely to give a birth estimate. The proportion of people affected with leprosy also varies by both `age` and `birest` (figure 3.8); people within any agegroup were more likely to be affected if they did not know their exact year of birth than those who did. The effect appears stronger in younger agegroups and there is a significant interaction between the two ($p < 0.001$). Although, overall, women are significantly more likely to be affected by leprosy than men, this effect is small (OR=1.16, $p < 0.001$) and is not significant after adjusting for `age` and `birest`.

To help decide how best to model the interaction, the logistic model with full interaction

$$Y \sim \texttt{age} + \texttt{birest} + \texttt{age.birest}$$

was fitted. The odds ratios and 95% confidence intervals for `birest` within agegroup are shown in table 3.3. The odds ratios for those aged 25+ are fairly constant, and although higher for agegroup 15–19 and 20–24, they are not significantly different from the 25+ group at the 5% level. The significant difference is between those aged 0–14 and those 15+. We decided not to include all possible interactions in the model, but instead to fit `birest` with 3 levels: 0 if exact year known, 1 if estimated year and the person was aged 15+ and 2 if they gave an estimated year and were aged 0–14.

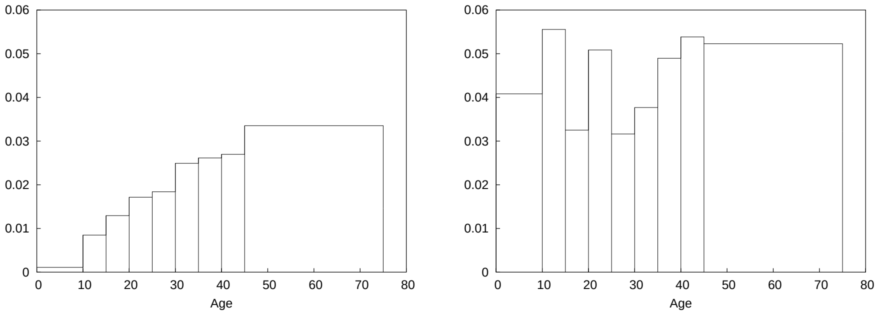
There is no evidence of interaction between any of the other covariates. The final definition of variables used in this and subsequent analyses is given in table 3.4 and the coefficients from fitting the logistic model

$$Y = 1 + \texttt{age} + \texttt{birest} + \texttt{pbcon} + \texttt{mbcon} + \texttt{bcg} + \texttt{sex}$$

are shown in table 3.5.

BCG scar status	Odds ratio for leprosy	95% CI
Before accounting for age:		
Positive	1.00	
Negative	2.60	(2.38, 2.86)
Uncertain	2.09	(1.84, 2.38)
After accounting for age:		
Positive	1.00	
Negative	1.41	(1.27, 1.58)
Uncertain	1.32	(1.15, 1.52)

Table 3.2: Odds of leprosy by BCG scar status



(a) Exact year of birth known (`birest`=0)

(b) Estimated year of birth given (`birest`=1)

Figure 3.8: Histograms showing cumulative incidence of leprosy, by the covariates `age` *and* `birest`

Age group	OR	95% CI
0–9	38.69	(9.18,163.06)
10–4	6.87	(2.75, 17.12)
15–9	2.56	(1.53, 4.28)
20–4	3.07	(2.26, 4.16)
25–9	1.74	(1.30, 2.32)
30–4	1.53	(1.18, 1.98)
35–9	1.91	(1.48, 2.46)
40–4	2.05	(1.59, 2.64)
45–74	1.59	(1.39, 1.81)
75–89	2.24	(1.10, 4.54)

Table 3.3: Odds ratios for disease comparing those giving an estimated year of birth to those giving an exact year of birth, by age group

Variable	Definition
`age`	broad agegroup at last examination before the end of LEP2
`birest`	categorical variable. 0 if exact year of birth known; 1 if estimated year given and the person was aged 15+; 2 if estimated year given and person aged 0–14
`bcg`	binary variable. 1 if a BCG scar status was positive, 0 if scar status was negative or uncertain
`pbcon`	binary variable. 1 if individual shared a household during LEP1 or LEP2 with a PB case, 0 otherwise
`mbcon`	binary variable. 1 if individual shared a household during LEP1 or LEP2 with a MB case, 0 otherwise
`sex`	binary variable. 1 if female, 0 if male

Table 3.4: Definition of variables used in epidemiological analysis

Covariate	OR	SE	p	[95% CI]
agegroup 0–9	1.00			
agegroup 10–14	7.44	1.08	0.000	[5.59, 9.91]
agegroup 15–19	11.75	1.67	0.000	[8.89, 15.53]
agegroup 20–24	18.02	2.54	0.000	[13.66, 23.77]
agegroup 25–29	17.28	2.48	0.000	[13.04, 22.90]
agegroup 30–34	19.96	2.86	0.000	[15.07, 26.44]
agegroup 35–39	21.91	3.16	0.000	[16.50, 29.08]
agegroup 40–44	23.06	3.34	0.000	[17.35, 30.65]
agegroup 50–74	24.30	3.29	0.000	[18.63, 31.68]
birest=1	1.76	.08	0.000	[1.61, 1.94]
birest=2	8.24	3.30	0.000	[3.76, 18.06]
pbcon	2.05	.08	0.000	[1.89, 2.24]
mbcon	2.98	.23	0.000	[2.55, 3.48]
sex	0.89	.03	0.006	[.82, .96]
scar	0.63	.03	0.000	[.56, .70]

Table 3.5: Results from fitting full logistic model

3.6 Discussion

The logistic regression shows that the most significant effect is due to age, with cumulative leprosy incidence increasing with age. The odds ratio for disease for household contacts of MB cases (2.98) is higher than that for PB cases (2.09). This is expected since MB cases, with a higher bacillary load, should be more infectious than PB cases. BCG scar positive people are less likely to have disease (OR=0.63), reflecting the protective effect of BCG vaccination. People over 15 who gave an estimated year of birth are at significantly increased risk of disease (odds ratio = 1.42). However, the few children under 15 who gave an estimated year of birth are at substantially increased risk of disease (odds ratio = 9.88). This is likely to be because those few children who do not know their exact year of birth (162 out of 76,231) form a very special group. Leprosy is known to be a disease of poverty and these children probably belong to one of the most impoverished groups - uneducated, and quite possibly orphans - therefore being at a much higher risk of disease.

The results of this preliminary analysis are in broad agreement with earlier analyses of this dataset (Pönnighaus et al., 1994). They will be useful in the next chapter as a check that a more complicated model is fitting well.

CHAPTER 4.

ESTIMATING THE RELATIVE RECURRENCE RISK RATIO

4.1 Introduction

The definition of the relative recurrence risk ratio, λ_R, and standard methods for its estimation were discussed in section 2.3. Guo (2000) has shown that ignoring environmental factors which influence susceptibility to disease can inflate estimates of this parameter.

Equation (2.1) may also be written

$$\lambda_R = \frac{P(D_1 = 1|D_2 = 1)}{P(D_1 = 1)} = \frac{P(D_1 = 1 \text{ and } D_2 = 1)}{P(D_1 = 1)P(D_2 = 1)}.$$

where $D_i = 1$ if individual i is affected and $D_i = 0$ otherwise. If non-genetic factors could be measured and the probabilities made conditional on these factors, any residual increase in risk would be

$$\lambda_R^* = \frac{P(D_1 = 1|D_2 = 1, X_1, X_2)}{P(D_1 = 1|X_1, X_2)} = \frac{P(D_1 = 1 \text{ and } D_2 = 1|X_1, X_2)}{P(D_1 = 1|X_1, X_2)P(D_2 = 1|X_1, X_2)} \quad (4.1)$$

where X_1 and X_2 denote covariate vectors. This residual risk would be due to unmeasured risk factors shared between relatives. If all non-genetic covariates are measured, these unmeasured factors must be genetic and λ_R^* must be the 'genetic relative recurrence risk ratio'.

This can be simplified further if we assume $D_1|X_1$ and $D_2|X_2$ are independent of X_2 and X_1 respectively, ie, conditional on each individual's own covariate data, their disease status does not depend on their relative's data. Then,

$$\lambda_R^* = \frac{P(D_1 = 1|D_2 = 1, X_1, X_2)}{P(D_1 = 1|X_2)} = \frac{P(D_1 = 1 \text{ and } D_2 = 1|X_1, X_2)}{P(D_1 = 1|X_1)P(D_2 = 1|X_2)}. \quad (4.2)$$

Note that there are cases where such an assumption may not hold - these are addressed in section 4.6.5 - but where it does, the attraction of the assumption is clear: λ_R is the ratio of the joint risk of disease in a pair of relatives to the product of their marginal risks, where the marginal risks are equivalent to those which would be estimated from a univariate population analysis of disease risk. (In such an analysis, it would be standard to model an individual's disease risk dependent on their covariates alone, ignoring any relatives' covariates, partly because there is no clear choice which relatives' covariates would be appropriate to include). This independence assumption allows that X_1 and X_2 may share some elements. For example, when analysing sibpair data, it may be appropriate to include some family-level covariate, eg employment status of household head as some proxy for socioeconomic status.

The definition of λ_R implicitly assumes individuals are either affected or not - ie it applies to binary traits which are manifest at birth, or at least soon afterwards. But most diseases are not like that; onset of disease is an event in time and for such disease it is necessary to consider familial aggregation in terms of association between times to disease onset. Note also that age at onset of disease is often of particular interest in genetic studies when researchers may believe that 'genetic cases' tend to have earlier onset than 'non-genetic cases'. In many studies it is possible to record time of onset; these are incidence studies. This is not always possible, however: in the KPS data, information about time of onset was considered unreliable (see section 3.1.2) and so we have only present state, or prevalence data. In either case, copula functions allow specification of models for association between disease state and onset times between individuals, and in this chapter a copula model which allows estimation of λ_R^* is developed and applied to the KPS leprosy data.

In the next section, copula functions are introduced, their statistical properties described and examples of their use in genetic epidemiology given. In section 4.3, a broad class of models, marginal models, are discussed, and their use in studies of familial aggregation of reviewed. Sections 4.4 and 4.5 describe the formulation of and method for fitting a marginal model which makes use of a particular copula function to estimate λ_R^*. This model is applied to leprosy data from the KPS and the results presented in section 4.6 while issues regarding its extension beyond relative pairs are discussed in section 4.7. Finally, a discussion of the use of the model developed, its relation to other models and interpretation of the results of the application to the KPS data is presented in section 4.8.

For the rest of this chapter, I will write λ_R for λ^*_R.

4.2 Copulas

Copulas are functions that join ('couple') multivariate distributions to their one dimensional margins. They can also be thought of as multivariate distribution functions with margins that are uniform on $(0, 1)$.

Sklar's theorem

Consider X and Y with distribution functions $F(x)$ and $G(y)$ and joint distribution function $H(x, y)$. There exists a unique copula C such that for all $(x, y) \in [\text{Range}F] \times [\text{Range}G]$ (which is $[0, 1] \times [0, 1]$ when F and G are continuous)

$$H(x, y) = C(F(x), G(y)).$$

For a proof, see Nelsen (1999), pp 15–18.

A copula function $C(F(x), G(y))$ describes the mapping of $(F(x), G(y))$ onto $H(x, y)$. Copulas are useful when the forms of the marginal distributions are known but the joint distribution is not, because they allow the creation of a joint distribution with given margins. In this application, copulas are used to examine the dependence between the leprosy disease status (affected/unaffected) of related individuals. For further theoretical details about copulas, see Nelsen (1999).

4.2.1 Properties of copulas

We first need two definitions:

Definition 4.1: Let S_1, S_2 be sets with least elements a_1, a_2 respectively. Then $H : S_1 \times S_2 \to \mathbf{R}$ is *grounded* if $H(x, a_2) = H(a_1, y) = 0 \ \forall \ (x, y) \in S_1 \times S_2$

Definition 4.2: Let $B = [x_1, x_2] \times [y_1, y_2]$ be a rectangle in the domain of H (Dom$H$). Then if

$$V_H(B) = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1)$$

H is *2-increasing* or *quasi-monotone* if $V_H(B) \geq 0 \ \forall \ B \in \text{Dom}H$.

Then we have

Definition 4.3: A *copula* is a 2-increasing, grounded function with domain $\mathbf{I}^2$ where $\mathbf{I} = [0, 1]$.

Equivalently, $C : \mathbf{I} \times \mathbf{I} \to \mathbf{I}$ such that

1. $\forall\, u, v \in \mathbf{I}$,
$$C(u,0) = C(0,v) = 0$$
$$C(u,1) = u; \quad C(1,v) = v$$

2. $\forall\, u_1 \leq u_2, v_1 \leq v_2 \in \mathbf{I}$
$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$$

4.2.2 Examples of copula models

Various classes of copula functions have been proposed and studied; examples include:

Clayton's copula (Clayton, 1978)

$$C_\theta(u,v) = \begin{cases} \left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}} & \theta > 0 \\ uv & \theta = 0 \end{cases}$$

Frank's copula (Frank, 1979)

$$C_\theta(u,v) = \begin{cases} -\frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right) & 0 < \theta < 1 \\ uv & \theta = 1 \end{cases}$$

Plackett's copula (Plackett, 1965)

$$C_\theta(u,v) = \begin{cases} \frac{1+(\theta-1)(u+v)-\sqrt{(1+(\theta-1)(u+v))^2-4uv\theta(\theta-1)}}{2(\theta-1)} & \theta \neq 1 \\ uv & \theta = 1 \end{cases}$$

4.2.3 Copulas in genetic segregation and linkage analysis

Meester and MacKay (1994) extended Frank's copula to deal with the multivariate case, and propose the copula

$$C_\alpha[F_1(y_1), \ldots, F_n(y_n)] = -\frac{1}{\alpha} log \left\{ 1 + (e^{-\alpha} - 1) \prod_{i=1}^{n} \left[\frac{e^{-\alpha F_i(y_i)} - 1}{e^{-\alpha} - 1} \right] \right\}$$

with

$$\lim_{\alpha \to 0} \{C_\alpha[F_1(y_1), \ldots, F_n(y_n)]\} = \prod_{i=1}^{n} F_i(y_i)$$

in the n-dimensional case. This is a one-parameter copula, with $-\infty < \alpha < \infty$ and $\alpha = 0$ denoting independence between the y_is. When $\alpha > 1$ or $\alpha < 1$, there is positive or negative dependence between the y_is respectively.

Trégouët et al. (1999) describe the use of this extended copula in a genetic context. They show how it may be used to calculate the probability of phenotypic data in a nuclear family, $P(\mathbf{y}|\mathbf{x}, \mathbf{g})$, where $\mathbf{y}$ denotes the vector of phenotypes, $\mathbf{x}$ the vector of covariates and $\mathbf{g}$ the vector of genotypes. This probability is necessary for segregation and joint segregation-linkage analysis. It is often likely there will be positive dependence between observations within a family; this means the joint probability is not a simple product of the marginal probabilities and the extended copula above may be used to calculate the joint probability with dependence accounted for by allowing $\alpha > 0$. Trégouët et al. (1999) demonstrate such an application in a combined segregation-linkage analysis of a binary trait, using high plasma levels of angiotensin converting enzyme (ACE) as an example. Their method is particularly useful because it will allow nuclear families of any size to be combined simply in one analysis. Extension to larger pedigrees is theoretically possible, but complicated because a further interaction parameter is needed for each additional relative type.

Li and Huang (1998) use the conditional hazard ratio to model the increase in risk to an individual due to sharing at least one allele IBD with an affected relative.

$$\theta(t_j, t_k) = \frac{\lambda_{T_j|T_k}(t_j|T_k = t_k)}{\lambda_{T_j|T_k}(t_j|T_k > t_k)}$$

for relatives j and k observed at times t_j and t_k and who become affected at times T_j and T_k respectively. For each individual, they construct neighbour sets who share one or other of the individual's alleles IBD at a locus (relatives who share both alleles are randomly assigned to one set or the other). Clayton's copula is used to model the joint survival function among each set of relatives and among the union of these sets, with the between and within sets conditional hazard ratios given by θ_0 and θ_1. They show how a test for linkage may be constructed as a pseudolikelihood ratio test of

$$H_0 : \theta_1 - \theta_0 = 0, \theta_1 > 1\_vs\_H_a : \theta_1 - \theta_0 > 0, \theta_1 > 1$$

where θ_1 is a nuisance parameter.

Pseudolikelihood has been used when the true likelihood is difficult to evaluate. Besag (1975, 1997) considered a situation where the likelihood of each of n random variables could be expressed conditional on its 'neighbours' (a subset of the other $n - 1$ variables) and proposed estimating the parameters of these conditional likelihoods by maximising the pseudolikelihood, defined as the product of the conditional likelihoods. In the case of Li and Huang, the likelihood of observed data for each neighbour set may be expressed in terms of their joint survival function, but since each member of a family appears in exactly two neighbour sets, it is difficult to write down the likelihood. Instead, the pseudolikelihood is defined as the product of likelihoods for each neighbour set within a family. The pseudolikelihood ratio test is then constructed as for a likelihood ratio, and is demonstrated to be asymptotically distributed as a 50:50 mixture of a (scaled) χ^2_1 and a degenerate at 0.

Their method is applicable to large numbers of pedigrees of moderate size. Sibling pairs are not suitable since the neighbour sets must either be null or contain only one member. The use of a copula to measure the joint survival function allows for correlation between age of onset in disease data between relatives not due to IBD sharing at the locus under test, though it offers no way to deal with cases where IBD sharing cannot be unambiguously determined and power calculations for this method have yet to be published.

4.3 Marginal models used in the study of familial aggregation of disease

In order to estimate λ_R, data on disease-status must be collected on sets of relatives, possibly with covariate data such as age, sex, known risk factors for disease and age at onset (if applicable). Often such data is ascertained through probands, sometimes only affected probands but also through cases and (possibly matched) controls. This is known as a family case-control design. Alternatively, ascertainment may be through sampling families (more often households) or by undertaking a complete population survey in a particular area (as in the KPS).

Once such data have been collected, the aim of the analysis may be

1. to infer relationships between covariates and disease in family members (accounting for correlation of disease and covariates within families), and/or
2. to evaluate disease correlations among family members, conditional on their covariates.

Log-linear models are commonly used in longitudinal studies to analyse repeated measures data, when multiple observations are recorded for the same individuals at different times and are therefore likely to be correlated. Familial data can also be considered as repeated measures data: multiple observations are made for each family, and are likely to be correlated due to shared genetic and/or environmental risk factors. Although log-linear models would therefore be applicable to such data, the natural parameters are in terms of conditional probabilities, which make the first aim above difficult to achieve.

Another method uses 'marginal multivariate regression models', or 'marginal models' (Liang et al., 1992). The joint distribution is parameterised in terms of marginal rather than fully conditional distributions of increasing order. For bivariate binary response data (Y_1, Y_2), the natural parameters are generally the mean response ($P(Y_1) = 1$, $P(Y_2) = 1$, which may be equal or unequal and may depend on covariates) and the odds ratio. A copula function may be used to relate the joint distribution function to the margins.

Often, relative trios and above are recruited as well as relative pairs. The likelihood can then be defined in terms of the joint likelihood of observing the relative set or in terms of the joint likelihood of observing each pair of relatives within the set. In the second approach, care must be taken when estimating standard errors.

Examples of marginal models which have been used to explore the familial aggregation of disease (some of which employ copulas and some of which do not) are discussed below.

4.3.1 Logistic regression of family data

Whittemore (1995) proposes a method for analysis of case-control family data using marginal models. Let $Y = (y_1, \dots, y_m)$ and $Z = (z_1, \dots, z_m)$ denote the disease status and covariate vectors for each proband (1) and their relatives $(2, \dots, m)$. Whittemore shows that data from family case-control studies may be treated as if collected under a prospective study with two separate samples, one from $P(Y_{-1}, Z|y_1 = 1)$ and one from $P(Y_{-1}, Z|y_1 = 0)$ (where the -1 suffix denotes the vector with the first (proband) element removed). This is possible when a specific independence assumption,

$$P(y_i = 1|Z) = P(y_i = 1|z_i),$$

holds, which implies that the covariates Z are independent of any unobserved source of disease correlation. This may often hold in genetic studies, where the gene influencing the disease is not thought to also influence other covariates, but care must be taken to test this assumption. Logistic margins are used, so

$$p_i = P(y_i = 1|z_i) = \frac{e^{\alpha+\beta z_i}}{1 + e^{\alpha+\beta z_i}}.$$

gives the form of the joint likelihood (with the intercept parameter, α, allowed to vary between probands and relatives) which may be maximised for a general $P(Y|Z)$.

To construct the joint likelihood for an example dataset consisting of ovarian cancer cases and controls and their mothers ($m = 2$), a class of models of the form

$$P(Y|Z) = \left(\prod_{i=1}^{2} p_i^{y_i} (1 - p_i)^{1-y_i} \right) (1 + \rho t_1 t_2) \tag{4.3}$$

are considered, where $t_i = (y_i - p_i)(p_i(1 - p_i))^{-1/2}$ is the ith standardised response, $i = 1, 2$. Treating ρ as a nuisance parameter and examining the estimated β allows the first aim to be investigated and treating β as a nuisance parameter and testing $\rho = 0$ allows the second aim to be investigated.

This model allows the marginal response among probands to differ from that of their relatives and takes account of the case-control design. Methods that do not take account of this may produce inconsistent estimates of odds ratios, although the independence assumption given by (4.3) needs to be met for results to be valid.

4.3.2 Bivariate discrete survival distribution

Shih (1998) proposed a bivariate discrete survival distribution which can accommodate covariates in the margins and yields a constant odds ratio at any grid point. They consider the situation when repeated bivariate observations are made at times $\{(l_1, l_2), \_l_1 = 1, \ldots, m_1, \_l_2 = 1, \ldots, m_2\}$. Suppose the bivariate failure times are given by (T_1, T_2) and let

$$p_j^l = P(T_j = l | T_j > l - 1), \quad j = 1, 2.$$

Then the (discrete) marginal survival function is

$$S_j(l) = P(T_j > l) = \prod_{k=1}^{l} (1 - p_j^k)$$

and covariates are included in the margins by modelling

$$p_j^l = h^{-1}(\mathbf{x}_{jl}' \boldsymbol{\beta})$$

where h is the logit link function (McCullagh and Nelder, 1983) and $\mathbf{x}_{jl}$ is a vector of covariate data for individual j measured at time l.

To model the association between T_1 and T_2, $\theta(l_1, l_2)$ is defined to be the odds ratio for failure at time (l_1, l_2) conditional on $T_1 > l_1 - 1$ and $T_2 > l_2 - 1$. For fixed $\theta(l_1, l_2) = \theta$ for all (l_1, l_2), the joint survival function may then be defined in terms of $S_j,\ j = 1, 2$ and θ. Covariates may also be incorporated into the joint survival function using a log-link:

$$\theta(l_1, l_2) = \exp\{\mathbf{z}_{l_1 l_2}' \boldsymbol{\alpha}\}.$$

This model is fitted using a two-stage procedure. The margins are fitted assuming independence of the bivariate failure times and $\boldsymbol{\alpha}$ is estimated by maximising a pseudo-likelihood ratio function with $\hat{\boldsymbol{\beta}}$ found in the first stage substituted for $\boldsymbol{\beta}$.

A possible extension to multivariate data is discussed, using the same margins, and pairwise odds ratios and conditional odds ratios to model the associations. This was applied to data on heart disease in sisters from a longitudinal study and produced an estimate of $\theta = 4.1$ for death from coronary heart disease (although calculation of any recurrence risk ratio was not discussed). This is a marginal model, with discretized hazard function margins, and the association measured using odds ratios. Although a copula function was not used to model the association, the model is similar to the one which will be proposed here (see below) and the similarities and differences will be discussed in section 4.8.

4.3.3 Lung cancer study

Schwartz et al. (1996) recruited 314 non-smoking lung cancer cases and 345 controls in Detroit, USA between 1984 and 1987; all were interviewed (when cases had died, a proxy, generally spouse, sibling, offspring or parent was interviewed) by telephone and questionnaire data was obtained from 2,252 and 2,408 family members of cases and controls respectively. These data consisted of medical histories, lung and other cancer incidence and death and covariates such as age, sex, smoking history and occupation and were analysed using two methods:

1. logistic regression to determine whether cases were more likely than controls to report that they had a first degree relative with lung cancer, adjusting for the case or control covariates;
2. logistic regression to model whether lung cancer status in relatives, adjusted for covariates, was associated with their status as a relative of a case or of a control.

Both analyses found significant evidence for increased risk of disease among relatives of cases only among those aged 40–59. The odds ratio for cases under 60 years reporting they had an affected relative (method 1) was 7.2 (95% CI 1.2–39.7) while the relative risk associated with being a relative of a case who developed disease before the age of 60 was 6.1 (95% CI 1.1–33.4). No significant effects were found for cases or relatives of cases aged 60+.

These data were reanalysed by Li et al. (1998) using the framework proposed by Whittemore (1995) and described above. The logistic margins were replaced with proportional hazards functions

$$\lambda(t|z) = \lambda_0(t)e^{\beta z}$$

to accommodate time at onset or censoring data and Clayton's copula was used to construct a joint survival function for relative sets (each proband plus his or her relatives)

$$S(x_1, \ldots, x_l) = \left[\sum_{j=1}^{l} S_j(x_j)^{1-\theta} - (l-1) \right]^{\frac{1}{1-\theta}}$$

where $S_j(x_j)$ is the marginal survival function for member j. The likelihood of the data is the product of the likelihood of the observed data among cases and controls and their families. The joint likelihood may be obtained through combination of this copula model and the proportional hazards margins to obtain estimates of θ, which is a measure of association between relatives.

The model was first fitted with no covariates which produced the estimate $\theta = 1.67$ (95% CI 1.46, 1.88). After accounting for covariates found to be significant in the original study, they found $\theta = 1.19$ (95% CI 1.00, 1.69). Note that $\theta = 1$ corresponds to no familial aggregation. Li et al. conclude there is little evidence for familial aggregation in lung cancer risk after environmental factors that may also aggregate in families are accounted for. However, they did not examine the relatives of the younger cases (aged 40–59) separately although they were found to be associated with increased familial risk in the first analysis. While there is no evidence for familial aggregation overall in either analysis, it is a common hypothesis that 'genetic' cases may have earlier onset and so it would have been interesting to see whether the copula model also found evidence for familial aggregation among relatives of the younger cases.

4.3.4 Washington Ashkenazi Study

In the Washington Ashkenazi Study (WAS), Wacholder et al. (1998) recruited over 5,300 volunteers from the Ashkenazi Jewish community in Washington who were genotyped for three specific BRCA1 and BRCA2 mutations known to increase carriers' risk of breast and ovarian cancer. They also answered a 20 minute questionnaire which focused on breast cancer risk factors in the volunteers and the vital status and history of cancer at several sites in their first and second degree relatives. Wacholder et al. refer to this as a ‘kin-cohort design’ to emphasise that the relatives of the probands formed a retrospective cohort followed from birth to onset of cancer or censoring time, but the term ‘genotyped proband design’ has also been proposed to emphasise that probands are genotyped and selected at random, conditional on disease status.

The penetrance of the mutations tested had been previously estimated in studies which recruited the relatives of cancer patients. Although such recruitment was used in order to obtain high frequencies of the mutations, Wacholder et al. argue that it may lead to overestimation of penetrance due to shared non-genetic risk factors in families. They used data collected in the kin-cohort study to estimate the risk of breast and ovarian cancer by age in carriers and non-carriers of the BRCA1 and BRCA2 mutations. Incidence was found to be higher among carriers, but that the penetrance of these mutations (cumulative incidence of disease to age 70) was lower (0.6–0.8) for each specific mutation than is reported in studies among relatives of breast and ovarian cancer patients (0.8–0.95).

The results of Wacholder et al. (1998) are similar to those found by Struewing et al. (1997) who also estimated penetrance of the BRCA1 and BRCA2 mutations by recruiting and genotyping Ashkenazi volunteers. They found the risk of breast cancer among carriers of any mutation to be 0.56 (95% CI 0.4–0.73) and of ovarian cancer to be 0.16 (95% CI 0.04–0.3).

Chatterjee et al. (2001) reanalysed the WAS data with the aim of examining the risk of disease among the first and second degree relatives of the volunteers, who had not been genotyped but whose history of cancer was obtained from questionnaires completed by the volunteers. They used a copula to model the joint survival function for breast cancer for relatives of the volunteers, and constructed the model as follows.

Let g^P denote the binary BRCA1/2 carrier status for a volunteer, and let g_i, $i = 1, \ldots, n$ denote the (unmeasured) carrier status for his or her n relatives who had age at onset T_i and were age t_i when the questionnaire was completed. Let δ_i denote their disease status and assume t_i to be independent of g_i and T_i.

For a pair of relatives (i, j) (not including the volunteer), let

$$P(T_i \geq t_i, T_j \geq t_j | g_i, g_j) = C_\theta(S_{g_i}(t_i), S_{g_j}(t_j))$$

where $S_1(t)$ and $S_0(t)$ denote the survival functions for carriers and non-carriers respectively. θ is then a measure of association between the marginal risks of disease after accounting for genotype. A quasi-likelihood of the observed data (disease status of relatives and carrier status of volunteer) was constructed by considering all pairs of relative for each participant and treating the pairs as if they were independent and partitioning on the carrier status of relatives:

$$\begin{aligned} L_{ij} &= P(T_i \geq t_i, T_j \geq t_j | g^P) \\ &= \sum_{g_i, g_j} P(T_i \geq t_i, T_j \geq t_j | g_i, g_j) P(g_i, g_j | g^P). \end{aligned}$$

A two-stage estimation procedure was used. First non-parametric margins were estimated and then the quasi-likelihood was fitted using the estimated $S_g(t)$, $g = 0, 1$ and Frank's, Clayton's and the Stable copula:

$$C_\theta(u, v) = \begin{cases} exp\left\{-\left[(-\ln u)^{\frac{1}{\theta}} + (-\ln v)^{\frac{1}{\theta}}\right]^\theta\right\} & 0 < \theta < 1 \\ uv & \theta = 1 \end{cases}.$$

They also demonstrate how the recurrence risk of disease in a woman conditional on disease in her relative and her carrier status may be calculated, which would be useful in genetic counselling. The recurrence risk ratio, defined as

$$\frac{\text{risk of disease}|\text{affected relative, carrier status}}{\text{risk of disease}|\text{no affected relative, carrier status}},$$

was higher (~ 2) among non carriers than carriers (~ 1.1, not significantly greater than one) and the the results from the different copula models were similar.

Chatterjee and Shih (2001) also proposed a bivariate cure-mixture model for the analysis of the same data. The mixture model measures two types of association - between age at onset (as above) and between overall lifetime susceptibility. This is done by categorising individuals as either susceptible or not. The pairwise odds ratio, γ, was used to measure association between overall disease susceptibility and formulate the joint survival distribution, conditional on both individuals being susceptible as a copula function of marginal survival distribution. The same three copula functions were used, and Chatterjee and Shih again found the results to be comparable between all three, providing evidence for a strong and significant association between overall susceptibility between pairs of relatives ($\gamma \simeq 2.8$). However, there was only weak evidence for association between age at onset once this association in overall susceptibility was accounted for.

4.4 Formulation of a model to estimate the relative recurrence risk ratio

The approaches described above all belong to the family of marginal models. When age at onset data is available, a hazard function is generally used to describe an individual's risk of disease at any given point in time and a survival function to describe the cumulative risk over time. The methods above which make use of copula functions do so to express the joint survival function in terms of marginal functions. But in some cases, age at onset data is not available - as discussed in section 3.1, the data from the KPS do not generally contain age at onset information. Instead, we may consider the data to be *present-state* or *point prevalence* data: we know the age at which each individual was observed and whether or not they had onset of clinical leprosy before that time. Generally, we do not know the age at which onset of disease occurred. In this case, an individual's cumulative risk of disease may be measured using cumulative prevalence.

A marginal model, is proposed here. Logistic margins are used to model each individuals risk of disease given measured non-genetic covariates. Plackett's copula (Plackett, 1965) is used to model the association between relatives using a pairwise odds ratio. This copula was chosen because it is the natural two-dimensional extension of the one-dimensional logistic function. This section introduces Plackett's copula and describes its extension to test hypotheses of interest in this study and the relationship of this extended copula to cross ratio models.

4.4.1 Plackett's copula

The odds ratio for the contingency table shown in table 4.1 is $\theta = \frac{ad}{bc}$. The

	+	-
+	a	b
-	c	d

Table 4.1: 2×2 *table*

value of θ describes the table as follows:

$$\theta \begin{cases} \in [0,1) & \text{observations are concentrated in } b,\ c \text{ cells} \\ = 1 & ad = bc, \text{ ie the cells are independent} \\ > 1 & \text{observations are concentrated in } a,\ d \text{ cells} \end{cases}$$

Plackett (1965) described a class of bivariate distributions with given margins and a single parameter to measure the degree of association. Let X, Y have marginal distribution functions $F_X(x), G_Y(y)$ and a joint distribution function $H_{X,Y}(x, y; \theta)$ (for simplicity denoted F, G and H) satisfying

$$\theta = \frac{H(1 - F - G + H)}{(F - H)(G - H)} \tag{4.4}$$

where θ is constant. Note that thus defined, θ can be interpreted as the odds ratio for the 2×2 table shown in figure 4.1.

For most H, θ will be a function of (x, y). But for some H, θ is a constant - such H are members of Plackett's family, also known as *constant global cross ratio* distributions or *contingency-type* (*C-type*) distributions.

Let $u = F(x), v = G(y)$ and C be the copula of X and Y. Then

$$\theta = \frac{C(u,v)[1 - u - v + C(u,v)]}{[u - C(u,v)][v - C(u,v)]}.$$

and

$$C_\theta(u,v) = \begin{cases} uv & \theta = 1 \\ \frac{1+(\theta-1)(u+v)\pm\sqrt{[1+(\theta-1)(u+v)]^2-4uv\theta(\theta-1)}}{2(\theta-1)} & \theta \neq 1 \end{cases} \tag{4.5}$$

For $\theta > 0$, $\theta \neq 1$, the root with a '+' sign is never a copula - that with a '-' sign always is (for a proof, see Nelsen (1999), pp. 80–1).

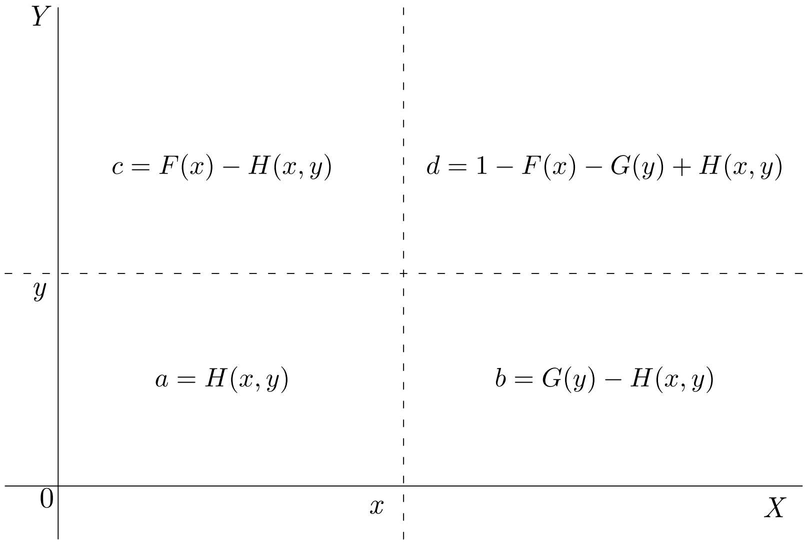


Figure 4.1: Contingency table for Plackett's copula

4.4.2 Extending Plackett's copula to estimate λ_R

In order to estimate λ_R, we need to be able to estimate the joint probability of affection for a pair of relatives. We can do this using the framework of a Plackett copula, which would require the odds ratio to be constant. For a disease such as leprosy, which is under control of genetic and non-genetic factors, it may not be realistic to assume θ is constant. For example, perhaps those who are genetically susceptible need only a low exposure to the infectious agent before they develop disease while those who are genetically resistant may need high exposure to develop disease. In such a case, θ would differ between high and low exposure groups. Plackett's copula is extended here; θ is not held constant, but expressed as a function of some joint covariates for the pair. Note this is similar to the method proposed by Shih (1998).

4.4.3 Relationship to cross ratio models

In fact, the above is a member of Dale's family of global cross ratio models (CRMs). Dale (1986) generalised Plackett's copula to ordinal bivariate data. Let y_{ij} represent the number of observations of the bivariate response $(Y_1, Y_2) = (i, j)$ - the data can then be summarised by the $r \times c$ contingency table shown in table 4.2. The table can be dichotomised along the double lines, and a series of $(r-1)(c-1)$ global odds ratios

$$\theta_{ij} = \frac{P(Y_1 \le i, Y_2 \le j)P(Y_1 > i, Y_2 > j)}{P(Y_1 \le i, Y_2 > j)P(Y_1 > i, Y_2 \le j)}, \quad (i \in 1, \ldots, r-1; j \in 1, \ldots, c-1)$$

can be calculated. Dale allowed the marginal probabilities to depend on a covariate vector $\mathbf{x}$ (through a logit link) and θ_{ij} to depend both on $\mathbf{x}$ and the cut point (i, j) (through a log link). This is a departure from the Plackett model which requires that the odds ratio be constant no matter where the dichotomisation occurs, and is therefore sometimes known as the constant global CRM. It has been applied to repeated measures data to allow for association between observations (e.g. Molenberghs et al., 1997).

4.5 Fitting the model using maximum likelihood

4.5.1 Estimating θ

Plackett avoids estimation of θ by maximum likelihood because it would be 'a tedious numerical process' and suggests the frequency estimate

$$\theta^{+} = ad/bc$$

where a, b, c, d are the observed frequencies given by lines in $\mathbf{R}^2$ parallel to axes through the point (p, q). θ^{+} is asymptotically normal, with mean θ and a variance estimated consistently by

$$V(\theta^{+}) = (\theta^{+})^2(1/a + 1/b + 1/c + 1/d).$$

y_{11}	$\cdots$	y_{1j}	$y_{1,j+1}$	$\cdots$	y_{1c}
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_{i1}	$\cdots$	y_{ij}	$y_{i,j+1}$	$\cdots$	y_{ic}
$y_{i+1,1}$	$\cdots$	$y_{i+1,j}$	$y_{i+1,j+1}$	$\cdots$	$y_{i+1,c}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_{r1}	$\cdots$	y_{rj}	$y_{r,j+1}$	$\cdots$	y_{rc}

Table 4.2: $r \times c$ contingency table for Dale's Cross Ratio Model (CRM)

The optimum choice for (p, q) is the sample median vector which minimises the asymptotic variance of θ^*. In this case, $F(p) = G(q) = 1/2$ and

$$\theta^* = \frac{4m^2}{(1-2m)^2}$$

where m is the observed frequency of observations in which neither variable exceeds its median.

But this estimate is only applicable if θ is held constant. Further, in many cases we observe only the discrete realization of the distribution and cannot choose where to place (p, q) - we can do nothing to improve the variance of the estimate. Instead, maximum likelihood methods are used here.

Consider present state (also called point prevalence) binary bivariate data - eg cumulative incidence of disease for pairs of relatives. Let (T_1, T_2) be the time at which pair of relatives become affected by disease and let the pair be observed at time (t_1, t_2). Let (d_1, d_2) be the disease status for the pair such that

$$d_i = \begin{cases} 1 & T_i \leq t_i \\ 0 & T_i > t_i \end{cases}; \quad i = 1, 2.$$

Assume that the disease is such that if people were affected by disease before they were seen, they either remain affected or show signs of past disease. Then, in the absence of censoring,

$$\theta = \frac{P(T_1 \leq t_1, T_2 \leq t_2) P(T_1 > t_1, T_2 > t_2)}{P(T_1 \leq t_1, T_2 > t_2) P(T_1 > t_1, T_2 \leq t_2)}. \tag{4.6}$$

Let $u_i = P(T_i \leq t_i), \ i = 1, 2$ be the marginal distribution functions for the probability of disease for two related individuals. Choosing a logistic link function for the margins, we have

$$\log \frac{u_i}{1 - u_i} = \boldsymbol{\beta} \mathbf{x}_i = \eta_i; \quad i = 1, 2$$

where $\mathbf{x}_i$ is a vector of covariates for individual i. Assume the joint distribution function $\delta = H(t_1, t_2; \theta)$ satisfies equation (4.4). Since $\theta \in [0, \infty)$, another natural link function is

$$\log \theta = \boldsymbol{\gamma} \mathbf{z} = \nu \tag{4.7}$$

where $\mathbf{z}$ denotes a joint covariate vector for the pair. Note that $\mathbf{z}$ need not contain any covariates, and setting $\mathbf{z} = 1$ gives Plackett's copula. Solving equation (4.4) gives the probability that both members of a pair are affected before time (t_1, t_2) as $\delta = C_\theta(u_1, u_2)$ as given by (4.5).

This model can be fitted using standard maximum likelihood theory. The log likelihood of such an observation is

$$\begin{aligned} L(d_1, d_2; \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) = {} & d_1 d_2 \log \delta + (1 - d_1) d_2 \log(u_1 - \delta) + \\ & d_1(1 - d_2) \log(u_2 - \delta) + (1 - d_1)(1 - d_2) \log(1 - u_1 - u_2 + \delta) \end{aligned}$$

and can be summed over all observed relative pairs to find the log-likelihood of all the observed data.

The likelihood could be maximised in two ways.

1. Two-stage procedure: first fit the marginals (either to the entire dataset or just those members of the relative pairs under consideration). Then use this model to get fitted values $(\hat{u}_1, \hat{u}_2)$ and fit the joint likelihood using these (see Shih and Louis, 1995, for further discussion of two-stage procedures in relation to copula models). This is simple, but the standard errors will be under-estimated in the fitted joint likelihood.

2. One-stage procedure: fit the joint likelihood and marginals together, maximising the joint likelihood with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ simultaneously. Standard errors can then be correctly estimated.

The first method has the advantage that the model for the marginals can be fitted using all individuals in the database and is therefore likely to be more accurate. The same model is then applied whichever relatives are used to fit the copula. The disadvantage, though, is that parameter estimates in the joint distribution (in particular, $\boldsymbol{\gamma}$ and hence θ and λ_R) will have inaccurate standard errors because the likelihood will have been maximised over marginal probabilities that are themselves estimates and not observed values. The second method has been used here because accurate estimates of θ and λ_R are the main goal.

Once maximum likelihood estimates of θ have been found, reference to equation (4.2) shows λ_R can be estimated by

$$\hat{\lambda}_R = \frac{\hat{\delta}}{\hat{u}_1 \hat{u}_2}.$$

4.5.2 Censoring

In both incidence and prevalence studies, removal of subjects from the study population before (or, for present state studies, after) disease onset, termed *censoring*, has the potential to distort findings. In the case of prevalence data, $P(T \leq t)$ is the cumulative incidence of disease, but we observe only prevalence which is conditional on the availability of subjects for study. Thus, we observe disease not with probability $P(T \leq t)$ directly, but with probability $P(T \leq t|C > t)$, where C is the censoring time, such that $C > t$ means that the subject remains in the study at time t. Appendix A formally explores what assumptions must be made about the censoring process if censoring is not to distort the results of any analysis.

4.5.3 Standard errors and multiple pairs

For a rare disease, there will be substantially more doubly unaffected relative pairs than pairs with one or both members affected. In this situation, it may speed analysis to take a sample of the doubly unaffected pairs and use weights accordingly when calculating the likelihood. Also note that not all pairs will be independent. A sibling trio can be split into three pairs, but the affected status of the third pair is completely determined by the status of the first two pairs. However, including only non-independent pairs could introduce a bias (a sib trio with two affected members could form two affected/unaffected pairs, or one affected/unaffected and one affected/affected pair).

An alternative to splitting trios and above into pairs would be to calculate the likelihood for the trio itself. However, Plackett's copula does not extend easily to more than two dimensions (Molenberghs and Lesaffre, 1994) and preliminary investigations in this project have shown that what might be expected to be natural parameterisations of the three dimensional model do not lead to valid copula functions (see section 4.7).

When using sampling, or non-independent observations, standard errors may be inaccurate. To take account of these issues, robust estimates of standard errors are used. This allows the relaxation of assumptions about the independence of observations. In particular, clustered robust estimates are used here, which allow for observations from the same family to be correlated (see Huber (1967); White (1982, 1980) and Royall (1986) for more detail).

4.5.4 Confidence Intervals

Estimating confidence intervals for the fitted values $\hat{u}_1, \hat{u}_2, \hat{\theta}$ is straightforward, since these are simply transformations of the estimated model parameters $\hat{\eta}_1, \hat{\eta}_2, \hat{\gamma}$. It is not so easy to calculate standard errors and confidence intervals for λ_R and δ because they are functions of more than one non-independent parameters. Instead simulation is used, but as this is time consuming, it is advisable to proceed to this step only after the best fit model has been determined.

After fitting the model, we have a variance-covariance matrix for all the parameters from which the variance covariance matrix Σ can be calculated for $\mathbf{x} = (\eta_1, \eta_2, \nu)'$. The expected values, $\boldsymbol{\mu} = (\hat{\eta}_1, \hat{\eta}_2, \hat{\nu})'$ are already known. The Cholesky decomposition of Σ can then be used to simulate trivariate normals from $N(\boldsymbol{\mu}, \Sigma)$ and for each realization $\mathbf{x}_i$ of $\mathbf{x}$, calculate $\mathbf{y}_i = (\lambda_R, \delta)'$. The empirical 95% confidence interval is then given by the 2.5% and 97.5% centiles of $\mathbf{y}$.

4.5.5 Interpretation of θ

θ can be thought of as the ratio of the odds of disease, given someone has a relative of type R with disease, to the odds of disease, given they have a relative (of type R) who is not affected. It is easier to make inferences about θ than λ_R because θ is a parameter in our model, and thus can take only a limited number of values, while λ_R can only be a fitted value and so will vary according to an individual's marginal probability of disease. Note that this makes sense - if susceptibility to a disease is affected by environmental factors, then the relative risk of disease will vary according to those factors. Also, $\hat{\lambda}_R$ is limited by $\hat{\theta}$ (see Appendix B) and will approach θ for rare diseases, so

$$\begin{cases} \hat{\lambda}_R \in [\hat{\theta}, 1) & \hat{\theta} < 1 \\ \hat{\lambda}_R = 1 & \hat{\theta} = 1 \\ \hat{\lambda}_R \in (1, \hat{\theta}] & \hat{\theta} > 1 \end{cases} .$$

However, estimates of λ_R for example individuals with specific covariates are easier to interpret and should also be reported. Plackett required that

θ is constant, while Dale allowed it to vary. Consider a disease that is influenced partly by genetics and partly by environment and suppose two genetic types exist in the population - susceptible and resistant. If both types are susceptible to disease at high environmental risk levels, but only genetically susceptible individuals are susceptible to disease even at low environmental risk levels, we would expect θ to be higher amongst low exposure groups, since the affected members of these groups would be mostly genetically susceptible. This could be seen as genetic risk factors modifying the effect of non-genetic risk factors. In this situation it may be beneficial to prefer people who are affected despite low environmental risks for inclusion in genetic analysis studies. This is similar to the argument used by researchers who focus genetic studies on those who have particularly early onset of some disease. On the other hand, θ may be constant across levels of environmental risks - in this situation there is no clear preference about who should be included in genetic studies. These two situations can be distinguished if we use $\mathbf{z}$ from (4.7) to dichotomies pairs according to their non-genetic risks.

4.6 Application to leprosy data from the KPS

The collection of data by the KPS was described in section 3. In this section I collect all the assumptions made when formulating the copula and show that the KPS data meets the requirements when we consider cumulative leprosy incidence by age. To meet these assumptions, some data will be excluded from the analysis. The same exclusions discussed in sections 3.2 and 3.4 will be used and are only summarised in this chapter.

multiple observations Many individuals (51.4%) were seen more than once. Each observation provides more information (particularly if disease onset was between two observations), but repeated observations of the same individual cannot be accommodated in the model, because they will be correlated with one another. In the case where an individual was seen more than once, age is recorded as the earliest observation at which an individual was recorded as a clinical leprosy case, or, for those never found to be cases, the latest observation before the end of LEP2.

complete ascertainment This is not explicitly required, and while the model could be extended to accommodate sampled data if the sampling method was well defined, it is not necessary in this case. The coverage of the two population surveys is estimated to have been very high and, as discussed in section 3.2, data collected after the end of LEP2 will not be included in this analysis.

known age at observation When we consider cumulative leprosy incidence with age, we need to know the age of an individual at their most recent observation. However, as described in section 3.3, 22.8% of individuals do not give an estimated year of birth, which means their exact age cannot be known. This can be dealt with by random assignment of a birth year to someone who has given an estimated year of birth, also described in section 3.3.

nondecreasing marginal distribution functions Cumulative incidence of leprosy is non-decreasing up to age 74, as shown in figure 3.7 and only those individuals aged 74 or less when last seen by the KPS are included in this analysis.

4.6.1 Choice of marginal covariates

Marginal covariates were chosen with reference to the results from section 3.5. All covariates found to have a significant effect in that analysis (defined in table 3.4) were included in the margins of this model.

4.6.2 Description of relative pairs

A total of 168,845 individuals were available for analysis; 2,911 (1.7%) of whom were recorded as having, or having had, leprosy. All relative pairs up to third degree were identified; great grandparent-grandchild and great aunt/uncle-niece/nephew pairs were excluded from the analysis because there were very few doubly affected pairs (2 and 13, respectively). The numbers of other relative pairs by affection status is shown in table 4.3.

Since it is known that estimates of λ_R are inflated when non-genetic risk factors, particularly those that tend to cluster within families, are ignored (Guo, 2000), it is interesting to consider here the correlation of covariates between relative pairs. Of particular interest is household contact. Table 4.4 presents the correlation coefficient, ρ, for each covariate used in fitting the model and the proportion of all relative pairs who were observed to share a household with each other during LEP1 or LEP2.

Relative pair	Number affected			Total
	0	1	2	
1st degree				
siblings	189,620	5,122	207	194,949
parent - child	215,770	9,777	236	225,783
2nd degree				
half siblings	166,910	3,782	99	170,791
aunt/uncle - niece/nephew	250,320	9,461	112	259,893
gd parent - gd child	156,070	8,622	73	164,765
3rd degree				
cousins	678,620	16,225	230	695,075
half aunt/uncle - niece/nephew	369,940	11,423	128	381,491

Table 4.3: Number of relative pairs by affection status and relative type

Relative pair	Pairwise correlation coefficient (ρ) for						% sharing
	`sex`	`bcg`	`mbcon`	`pbcon`	`birest`	`age`	household
parent	.005	.016	.581	.614	.204	.196	77.9
siblings	-.003	.186	.604	.592	.372	.811	67.0
gd parent	-.005	.016	.239	.240	.061	.115	21.9
aunt/unc	.004	-.015	.145	.162	.018	.474	12.7
h. siblings	.018	.088	.485	.443	.003	.483	44.3
h. aunt/unc	.004	.025	.130	.122	.007	.362	7.1
cousins	-.003	.057	.081	.082	.118	.377	4.0

Table 4.4: Correlation between covariates used in estimation of λ_R *between relative pairs and proportion of relative pairs observed to share a household during LEP1 or LEP2*

4.6.3 Conditions for fitting of model

The copula model was fitted under the following sets of conditions:

1. no covariates
2. (a) all covariates excluding household contact; θ held constant
   (b) all covariates excluding household contact; θ allowed to vary
3. (a) all covariates; θ held constant
   (b) all covariates; θ allowed to vary

θ was allowed to vary under conditions 2(b) and 3(b) by fitting a separate θ for those pairs with high and low non-genetic risk factors. Individuals were dichotomised according to their marginal predicted risks (within relative types) under conditions 2(a) and 3(a) respectively. In part, this was to check whether the assumption under conditions 2(a) and 3(a) (that θ is constant) was valid. If it was not, it was hoped this would enable us to distinguish whether the genetic effect was stronger among those with lower or higher non-genetic risk factors. The cutoff chosen was the median marginal risk among affecteds estimated under conditions 2(a) and 3(a). It was hoped this would provide most power by placing equal numbers of affecteds in each group. Pairs were then categorised as low-low, low-high or high-high.

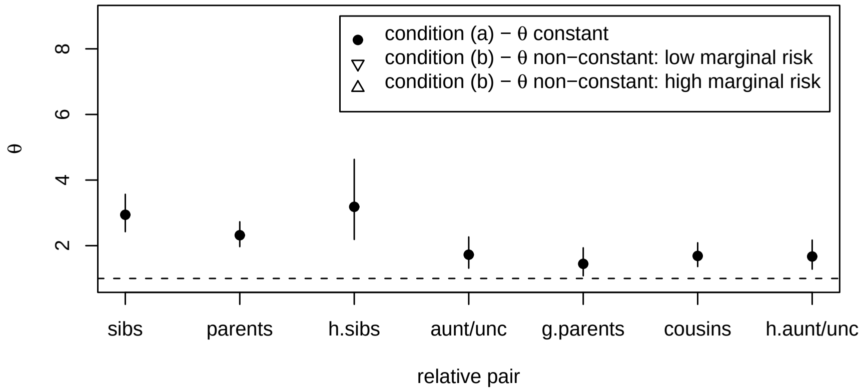
When fitting the model under conditions 2(b) and 3(b), no significant difference could be found between the low-high and high-high pairs, so the two groups were combined into a single high risk group.

4.6.4 Results

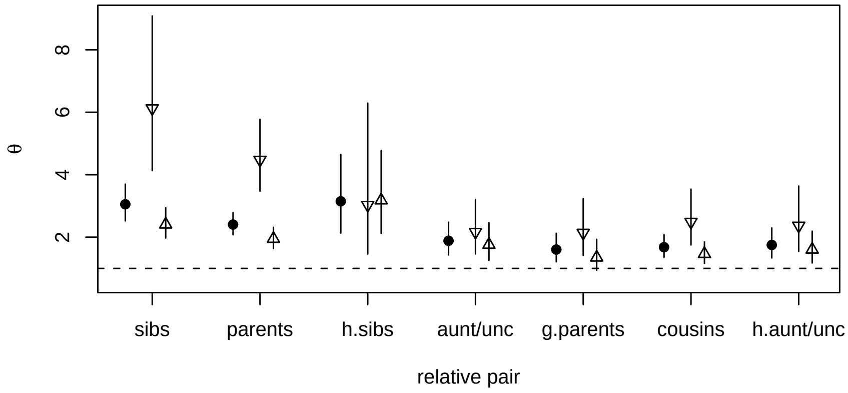
The results of fitting the cross-ratio model are presented initially in terms of θ. This is because θ is a parameter in the model while λ_R is a fitted value as discussed in section 4.5.5, so θ is more useful for discrimination between conditions.

Estimates of θ calculated from fitting the model under each condition are shown in figure 4.2. Under condition 2(b), θ varied significantly between low and high risk pairs, however, under condition 3(b) there was no evidence that θ varied for any relative pairs.

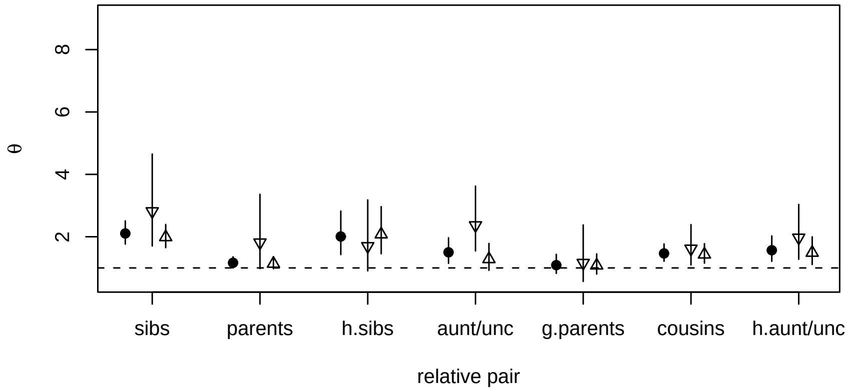
Fitting the CRM and ignoring covariates (condition 1) gives estimates of θ between 1.5 and 3.2 and hence λ_R. These do not change when sex



(a) Condition 1



(b) Condition 2



(c) Condition 3

Figure 4.2: Estimates of θ from fitting the cross ratio model (CRM) by relative type

and BCG covariates are included (condition 2(a)), but fall when household contact covariates are also included (condition 3(a)), reflecting the tendency for household contact but not sex or BCG status to cluster among relatives.

There is significant evidence under condition 2(b) that θ is not constant for most relative pairs when all covariates except household contact are included in the margins. This indicates that the model fitted under condition 2(a) does not explain all the variation in the data.

If we believed that all risk factors had been included in the margins (ie had we not been aware of or been able to measure household contact), then these results may have led us to conclude that the genetic effect varied at different levels of non-genetic risk. For example, younger pairs who were both leprosy cases might have been under a stronger genetic influence than older pairs.

Under condition 3(b), however, there is no evidence that θ may not be constant. This indicates that the model fitted under 3(a) could not be improved by relaxing the requirement that θ be held constant. It also indicates that the strength of genetic effect is constant across different levels of non-genetic risk. Conditions 3(a) appear to produce the best fit.

Had we not measured household contact, we might then have targeted future genetic studies at pairs predicted by the model to have higher θ (and hence λ_R) values. In fact, condition 3(b) shows θ does not vary, and targetting pairs in this way would have introduced more work with no likely benefit. This, together with the much higher estimates of λ_R observed under conditions 1 and 2, emphasises that while the model detects unmeasured shared risk, it is important that all non-genetic risk factors are measured if we hope our estimates to reflect genetic risk.

The preferred condition, then, is 3(a) and odds ratios from this model are given in table 4.5. Comparison with table 3.5 shows that the odds ratios for the marginal parameters are very similar, indicating that the margins of this copula model are fitting as expected. Histograms of the fitted λ_R values for each relative pair are shown in figure 4.3. λ_S is concentrated in the interval $[1.8, 2.0]$. It is interesting to note that λ_{HS} is in the same interval, which λ_O (parent-offspring) is only just above 1. Similarly, λ_{GP} (grandparent-grandchild) is just below 1, while other λ_R values are around 1.4. These results are discussed further in section 4.8.2.

Covariate	OR	p value	[95% CI]
Marginal model			
agegroup 0–9	1.00		
agegroup 10–14	7.43	$< 10^{-4}$	[6.07, 9.08]
agegroup 15–19	12.11	$< 10^{-4}$	[10.12, 14.50]
agegroup 20–24	17.88	$< 10^{-4}$	[14.76, 21.66]
agegroup 25–29	18.80	$< 10^{-4}$	[15.39, 22.96]
agegroup 30–34	23.21	$< 10^{-4}$	[18.96, 28.41]
agegroup 40–44	23.12	$< 10^{-4}$	[18.72, 28.56]
agegroup 45–74	23.82	$< 10^{-4}$	[19.83, 28.61]
birest=1	1.72	$< 10^{-4}$	[1.59, 1.87]
birest=2	4.04	$< 10^{-4}$	[2.08, 7.86]
pbcon	2.07	$< 10^{-4}$	[1.90, 2.25]
mbcon	2.64	$< 10^{-4}$	[2.24, 3.10]
scar	0.62	$< 10^{-4}$	[0.57, 0.68]
sex	0.91	< 0.01	[0.85, 0.98]
Joint model			
siblings	2.00	$< 10^{-4}$	[1.68, 2.39]
grand parents	0.98	0.92	[0.74, 1.31]
aunt/uncle	1.38	< 0.01	[1.05, 1.79]
half siblings	2.00	$< 10^{-4}$	[1.45, 2.75]
h aunt/uncle	1.48	$< 10^{-3}$	[1.16, 1.89]
cousins	1.43	$< 10^{-4}$	[1.18, 1.73]
parent	1.11	0.18	[0.94, 1.32]

Table 4.5: Parameter estimates from the model fitted under the preferred condition, 3(a)

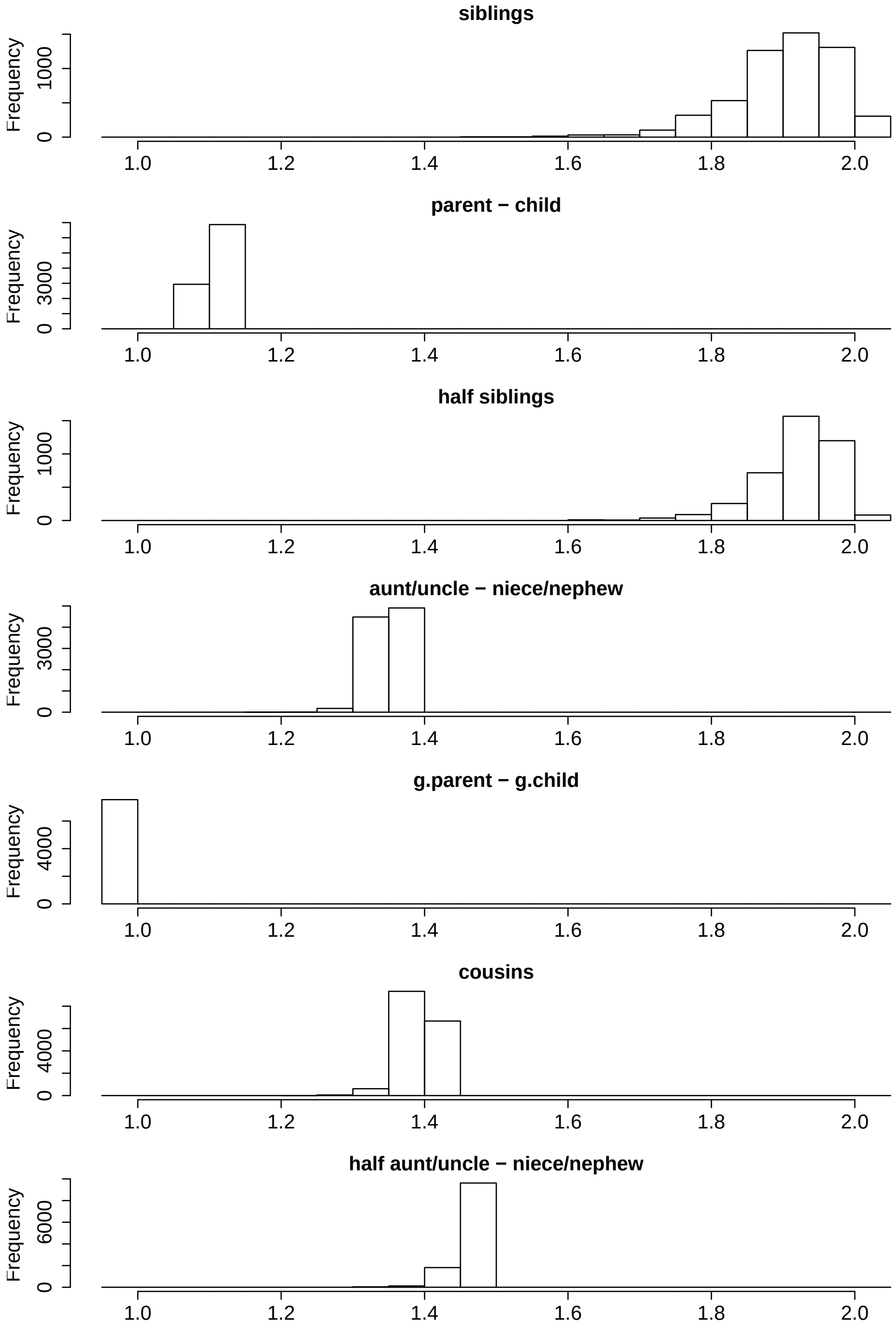


Figure 4.3: Frequency histograms of the fitted λ_R values under condition 3(a) by relative pair

4.6.5 Dependence of disease status on covariates of relatives

Equation (4.2) was derived from equation (4.1) under the assumption that $D_1|X_1$ is independent of X_2, and vice-versa. This section presents the arguments for that assumption to hold and also discusses whether it holds in the KPS dataset and possible bias which may result if it does not.

For the assumption to hold, we require $E(D_1|X_1, X_2) = E(D_1|X_1)$ (and similarly for D_2, but, without loss of generality, the focus of this section will be on D_1). In other words, once X_1 is measured, no further information about D_1 could be obtained from measuring X_2. If believe X_1 can be measured accurately and define X_1 to contain all covariates of interest for individual 1 and any family level covariates of interest for both individuals 1 and 2, it seems plausible that this requirement holds and so the independence assumption is met.

However, if measurements of (X_1, X_2) are made with error, so that we do not observe X_1 and X_2 directly, but W_1 and W_2, this may not hold. In this case, since X_1 and X_2 are likely to be correlated (see section 4.6.2), W_2 *will* contain further information about X_1 (and hence D_1) even after W_1 is known. When using categorical covariates (as we will here), this is often termed *misclassification.* Two types of measurement error are commonly distinguished:

- *non-differential error* where the error does not depend on disease status. Symbolically, this means $E(D_1|X_1, W_1) = E(D_1|X_1)$.
- *differential error* where the measurement error differs between those affected and non-affected individuals.

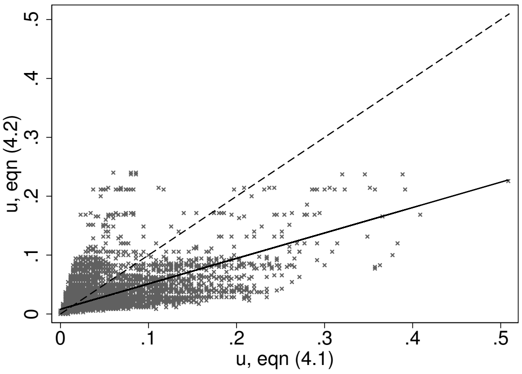
An example of differential error is recall bias, where affected individuals may have a different recollection of some exposure than unaffected individuals. In the collection of the KPS data, efforts were made at all stages to collect the most accurate covariate data possible on individuals, regardless of their disease status (which would often not be known by the interview teams). It is therefore unlikely that substantial differential error could be present in the KPS data, and we can restrict our attention to non-differential errors, which lead to a more predictable pattern of bias.

It is well known that non-differential errors in the measurement of covariates leads to *regression dilution* - the effects of covariates are underestimated, and the predicted risk of disease for individuals is over(under)estimated for those whose 'true risk' is below (above) the population mean (Carroll et al., 1995). It is likely that such errors would lead to similar over- and underestimation of the joint risk of disease, δ, but it is not clear in which direction this might bias estimates of λ_R - this would depend on the relative magnitude of bias in the joint and marginal predicted risks. It is even conceivable (though unlikely) that the bias in u and δ may cancel so that estimates of λ_R are unbiased.

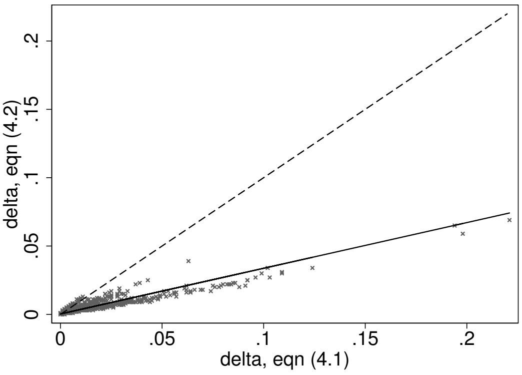
To check the effect of applying the independence assumption, the model was refitted using just the sibship data but using equation (4.1) to estimate u_1, u_2, δ, θ and λ_S. To distinguish these estimates from those calculated above (using equation (4.2)), they will be suffixed with †. Figure 4.4(a) shows a scatter plot of u vs $u^\dagger$, demonstrating that, as in the measurement error situation, estimates of u were attenuated. Since δ is a function of (u_1, u_2), a similar effect might be expected and is found for estimates of δ (figure 4.4(b)). This indicates that the independence assumption was not met, and that a worrying bias might have been introduced by incorrectly applying this assumption. However, estimates of θ and λ_S appear considerably less affected. We find $\hat{\theta} = 1.96$ and $\hat{\theta}^\dagger = 2.01$, with no significant difference, and figure 4.4(c) shows that λ_S and $\lambda_S^\dagger$ are closely correlated with only slight attenuation.

Since siblings' covariates are most closely correlated (table 4.4), it would be expected that any bias introduced by incorrectly applying the independence assumption would be strongest for these relatives. It is comforting then, that even in this case, when the bias in (u_1, u_2) and δ can be quite considerable, the bias in the estimates of λ_S is relatively small ($< 20\%$). This indicates that there is little bias in the results in this chapter (particularly not with respect to estimates of θ), despite incorrect application of an independence assumption.

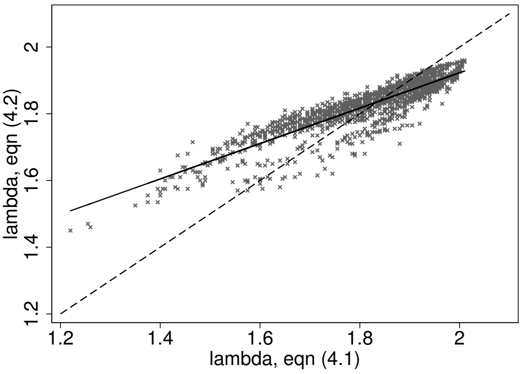
It is also clear from this section that the covariates used in this analysis are likely to have been measured with error, as acknowledged in chapter 3, but it is hard to speculate on just how large this error might be. However, we have shown in appendix A that estimates of θ are unbiased even when estimates of (u_1, u_2) may be biased due to censoring, and in this section that even quite considerable bias in estimates of (u_1, u_2) resulting from incorrect application of an independence assumption do not lead to significant bias



(a) u vs $u^\dagger$



(b) δ vs $\delta^\dagger$



(c) λ_S vs $\lambda_S^\dagger$

Figure 4.4: Bias in estimates of u, δ and λ_S resulting from the independence assumption in equation (4.2). The solid line shows the best fit regression line

in estimates of θ and only small bias in estimates of λ_R. Therefore, it is plausible to believe that estimates of θ and λ_R using this method are not strongly biased even in the presence of errors in variables.

4.7 Extension: modelling the association between trios of relatives

From the above, it appears that there may be a natural extension of the copula for modelling the association of disease state between trios of relatives and above. Molenberghs and Lesaffre (1994) proposed an extension of the CRM model to multivariate data. They noted that the bivariate cross ratio (θ, above), may be viewed as the ratio of two conditional odds ratios:

$$\theta_{12} = \frac{\theta_{1|2}}{\theta_{1|\bar{2}}}$$

where the numerator is the odds ratio for individual 1, conditional on disease in individual 2 and the denominator is conditional on no disease in individual 2. They used the marginal odds ratios and cross ratios to parameterise the margins and bivariate associations, and proposed modelling the trivariate association using the ratio of conditional bivariate odds ratios:

$$\psi_{123} = \frac{\theta_{12|3}}{\theta_{12|\bar{3}}}$$

(with the numerator and denominator defined similarly to above).

They showed how such a parameterisation may be generalised to n dimensions (taking ratios of conditional $(n-1)$ dimensional odds ratios to describe the n dimensional association) and further generalised to include ordinal data (the multivariate Dale model).

They also note that ‘the parameter space of the marginal odds ratios is constrained’ in the trivariate case and above and that ‘not every combination leads to a valid solution’. These constraints are not explored in any detail, and it is suggested the constraints are mild because they did not encounter problems maximising the likelihood for an example dataset.

In this section I explore both this and an alternative formulation for a trivariate extension and whether the criteria for a copula (see section 4.2.1) would be met. It is shown that they do not hold in general for either of two parameterisations considered.

4.7.1 Notation

We observe relative trios (containing members $(1, 2, 3)$) of type R at times (t_1, t_2, t_3). Let T_i be the time at which member i becomes affected by and showing clinical signs of leprosy and let

$$\begin{cases} u_i(t_i) & \text{denote } P(T_i \leq t_i), \quad i = 1, 2, 3 \\ u_{ij}(t_i, t_j) & \text{denote } P(T_i \leq t_i, T_j \leq t_j), \quad i, j = 1, 2, 3, \quad i \neq j \\ \delta(t_i, t_j, t_k) & \text{denote } P(T_i \leq t_i, T_j \leq t_j, T_k \leq t_k) \end{cases} \tag{4.8}$$

To simplify notation, the explicit dependence on time is again removed and we write $u_i(t_i) = u_i$ etc. Just as a 2×2 table was used in chapter 4, a $2 \times 2 \times 2$ table is used here, represented in figure 4.5. The events represented in each section are

$$\begin{cases} a = \text{P(1,2,3 affected)} & = \delta \\ b = \text{P(2,3 affected)} & = u_{23} - \delta \\ c = \text{P(1,3 affected)} & = u_{13} - \delta \\ d = \text{P(3 affected)} & = u_3 - u_{23} - u_{13} + \delta \\ e = \text{P(1,2 affected)} & = u_{12} - \delta \\ f = \text{P(2 affected)} & = u_2 - u_{23} - u_{12} + \delta \\ g = \text{P(1 affected)} & = u_1 - u_{13} - u_{13} + \delta \\ h = \text{P(none affected)} & = 1 - u_3 - u_2 - u_1 + u_{23} + u_{13} + u_{12} - \delta \end{cases}$$

There are thus 7 independent probabilities to consider ($h = 1 - a - b - c - d - e - f - g$).

4.7.2 Parameterisation of the extended model

We can parameterise the problem in two ways, using either marginal or conditional odds ratios. The marginal odds parameterisation is perhaps the most natural extension to the two dimensional case, but it was hoped the conditional parameterisation would lead to a more mathematically tractable solution.

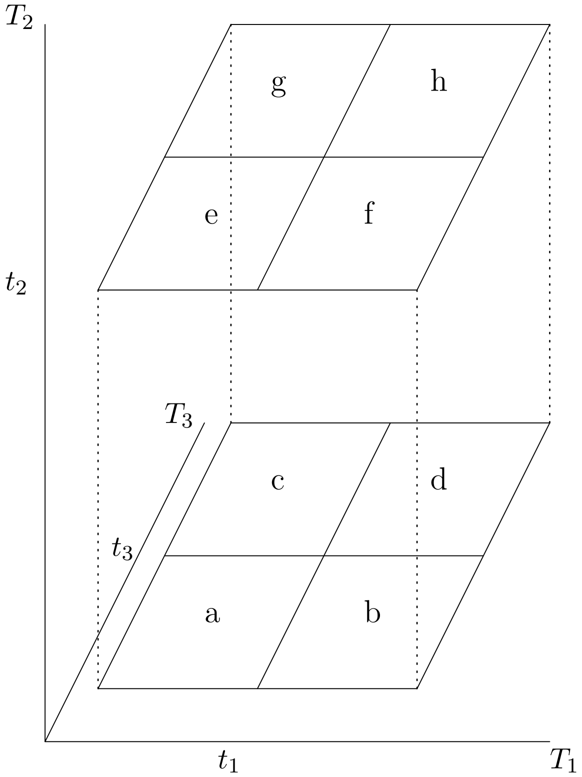


Figure 4.5: $2 \times 2 \times 2$ *table*

Parameterisation using marginal odds ratios

We use the marginal probabilities u_1, u_2, u_3, three marginal odds ratios

$$\theta_{12} = \frac{(d+h)(e+a)}{(b+f)(c+g)} = \frac{(1+u_{12}-u_2-u_1)u_{12}}{(u_2-u_{12})(u_1-u_{12})} \tag{4.9}$$

$$\theta_{13} = \frac{(f+h)(a+c)}{(b+d)(e+g)} = \frac{(1+u_{13}-u_3-u_1)u_{13}}{(u_3-u_{13})(u_1-u_{13})} \tag{4.10}$$

$$\theta_{23} = \frac{(g+h)(a+b)}{(c+d)(e+f)} = \frac{(1+u_{23}-u_3-u_2)u_{23}}{(u_3-u_{23})(u_2-u_{23})} \tag{4.11}$$

(where θ_{ij} is the marginal odds ratio for members i and j) and a conditional odds ratio

$$\begin{aligned}\psi &= \frac{ad/bc}{eh/fg} \\ &= \frac{\delta(u_3+\delta-u_{23}-u_{13})(u_1-u_{12}-u_{13}+\delta)(u_2+\delta-u_{23}-u_{12})}{(u_{23}-\delta)(u_{13}-\delta)(1-\delta+u_{23}+u_{13}-u_3+u_{12}-u_2-u_1)(u_{12}-\delta)}\end{aligned} \tag{4.12}$$

which is the ratio of the odds ratio for disease in relatives 1 and 2, conditional on disease in member 3 to the odds ratio for disease in relatives 1 and 2, conditional on no disease in member 3.

Assuming these odds ratios and marginal probabilities (u_i) can be estimated from data, the pairwise probabilities (u_{ij}) can be estimated as described in chapter 4 and the only unknown is δ, which is the solution to the quartic:

$$a_4 z^4 + a_3 z^3 + a_2 z^2 + a_1 z + a_0$$

where

$$\begin{aligned}
a_0 &= \psi u_{12} u_{23} u_{13} S \\
a_1 &= -[(u_{23}u_{13} + u_{12}u_{23} + u_{12}u_{13})S + u_{23}u_{13}u_{12}]\psi \\
&\qquad - (u_3 - u_{23} - u_{13})(u_1 - u_{12} - u_{13})(u_2 - u_{23} - u_{12}) \\
a_2 &= [u_{23}u_{13} + u_{12}u_{23} + u_{12}u_{13} + (u_{12} + u_{23} + u_{13})S]\psi \\
&\qquad\qquad - (u_3 - u_{23} - u_{13})(u_1 - u_{12} - u_{13}) \\
&\qquad\qquad - (u_3 - u_{23} - u_{13})(u_2 - u_{23} - u_{12}) \\
&\qquad\qquad\qquad - (u_1 - u_{12} - u_{13})(u_2 - u_{23} - u_{12}) \\
a_3 &= -1 - (u_{12} + u_{23} + u_{13} + S)(\psi - 1) \\
a4 &= \psi - 1 \\
S &= 1 + u_{12} + u_{23} + u_{13} - u_3 - u_2 - u_1
\end{aligned}$$

There are several algorithms for finding solutions to quartics. The general procedure is to substitute z so that the cubic term is removed, and using further substitution to reduce the problem to finding one root of the *resolvent cubic*. This root can then be subsitituted back and the solution to the original quartic easily found. Unfortunately there is no ‘nice’ algebraic form in which these roots may be expressed.

Parameterisation using conditional odds ratios

We use the marginal probabilities u_1, u_2, u_3, three conditional odds ratios

$$\begin{aligned}\theta_{12|3} &= \frac{af}{eb} = \frac{\delta(u_2+\delta-u_{23}-u_{12})}{(u_{12}-\delta)(u_{23}-\delta)}\\ \theta_{13|2} &= \frac{ad}{bc} = \frac{\delta(u_1+\delta-u_{12}-u_{13})}{(u_{12}-\delta)(u_{13}-\delta)}\\ \theta_{23|1} &= \frac{ag}{ec} = \frac{\delta(u_3+\delta-u_{13}-u_{23})}{(u_{13}-\delta)(u_{23}-\delta)} \qquad (4.13)\end{aligned}$$

(where $\theta_{ij|k}$ is the odds ratio for disease in members i and j, conditional on disease in member k) and the conditional odds ratio, ψ, as given in equation (4.12).

Again, assuming these odds ratios and marginal probabilities can be estimated from data, we now have four unknowns, the 3 pairwise probabilities (u_{12}, u_{13}, u_{23}) and δ. These can be found by simultaneous solution of the equations (4.12) and (4.7.2)–(4.13), but again, there is no algebraic solution.

4.7.3 Assessment of the criteria for a copula

To be a copula, the solution $\delta = C(u_1, u_2, u_3)$ of our quartic must satisfy certain relations. These are (from Nelsen, 1999, definition 2.10.6)

1. C is a function mapping $[0,1]^3 \to [0,1]$, such that

   $$\delta = C(u_1,u_2,u_3) = 0 \text{ if at least one of } u_1,u_2,u_3 \text{ is } 0$$

   and

   $$\text{if all } u_i \text{ are 1 except } u_k, \text{ then } C = u_k,\ k \neq i$$

2. C is *3-increasing*. That is, $V_C(B) \geq 0$ for all boxes B whose vertices lie in Dom C, where the volume of the box B $[u_1,u_1'] \times [u_2,u_2'] \times [u_3,u_3']$ is given by

   $$\begin{aligned}V_C(B) = {} & C(u_1',u_2',u_3') - C(u_1',u_2',u_3) - C(u_1',u_2,u_3') - C(u_1,u_2',u_3')\\ & + C(u_1',u_2,u_3) + C(u_1,u_2',u_3) + C(u_1,u_2,u_3') - C(u_1,u_2,u_3)\end{aligned}$$

Since neither of our parameterisations have algebraic solutions, the conditions are difficult to prove analytically. Instead, I will enumerate the solution for a five-dimensional grid of values. To simplify things, set the second order odds ratios (eg θ_{12} or $\theta_{12|3}$) equal to one another, since it has already been shown that the two dimensional odds ratio for leprosy in this population is constant (section 4.6.4). The grid used is

$$\begin{aligned} G = \{ & u_1 \in \{0.0001, 0.1, 0.2, \ldots, 0.8, 0.9, 0.999\}, \\ & u_2 \in \{0.0001, 0.1, 0.2, \ldots, 0.8, 0.9, 0.999\}, \\ & u_3 \in \{0.0001, 0.1, 0.2, \ldots, 0.8, 0.9, 0.999\}, \\ & \theta \in \{0.75, 1.00, 1.20, 2.0, 3.0, 5.0\}, \\ & \psi \in \{0.75, 1.00, 1.20, 2.0, 3.0, 5.0\}\} \\ \text{s.t. } & \theta \leq 1 \Longleftrightarrow \psi \leq 1 \quad \text{and} \quad \theta \geq 1 \Longleftrightarrow \psi \geq 1 \end{aligned}$$

We can then check that the properties of a copula are satisfied by testing whether

1. for any point in G, there exists a single solution which meets the conditions

   $$\begin{aligned} u_{12} &\in (0, \min\{u_1, u_2\}) \\ u_{13} &\in (0, \min\{u_1, u_3\}) \\ u_{23} &\in (0, \min\{u_2, u_3\}) \\ \delta &\in (0, \min\{u_{12}, u_{23}, u_{13}\}) \end{aligned}$$

   (which must hold because of the definition of these probabilities in equation (4.8))

2. for $u_1 \to 0$, $\delta \to 0$

3. for $u_1 = u_2 \to 1$, $\delta \to u_3$

4. for all boxes $B \in G$, $V_C(B) \geq 0$

Marginal odds ratios

There are several points in G where there are two distinct real roots for δ in the desired range. We do not have a unique solution, so this parameterisation does not lead to a copula.

Conditional odds ratios

Criteria 1–3 appear to be met. To test criteria 4, $V_C(B)$ was calculated for a total of nearly four million boxes across G with all parameters allowed to vary, subject to the restriction $\theta_{12|3} = \theta_{13|2} = \theta_{23|1}$. Of these, 7.2% had negative volume; the least volume was -0.667. This is not within what might be expected given rounding error, and shows that δ is not 3-increasing and so this parameterisation does not lead to a valid copula either. However, since all other conditions were met, it is possible that this parameterisation would lead to a well-defined copula should further suitable conditions be imposed, but there was not time in this project to explore this further.

4.8 Discussion

4.8.1 Proposed model

We have proposed a fully parametric marginal model with logistic margins and the joint distribution specified by an extended Plackett copula, from which λ_S fitted values may be found. This differs from previous approaches in this area mainly because of the different nature of the available data. Li et al. (1998) and Chatterjee et al. (2001) made use of case-control family data which included age at onset and used copula functions to specify the joint survival function. This was not possible in our application. The model proposed in this paper is similar to another marginal model approach, not involving copulas, proposed by Shih (1998). Data was available from a longitudinal survey and also modelled associations between observations from related individuals were modelled with a constant odds ratio. Again, repeated measures on individuals at regular time intervals were available (and attention was restricted to those seen 10 times or more). This allowed a discretized survival distribution to be fitted. However, Shih (1998) used a two-stage estimation procedure, first fitting the margins assuming independence and then substituting these fitted values into a pseudo-likelihood which was maximised. Our likelihood can be maximised simultaneously with respect to marginal and joint distribution parameters which allows standard errors of parameters to be accurately estimated.

The other difference lies the treatment of relative trios and above. The Li et al. (1998) framework models the joint survival distribution for the proband and all his/her relatives, while Chatterjee et al. (2001) considered only the joint survival distribution among pairs of relatives (not including the proband). Although Shih (1998) models associations in larger relative sets using conditional odds ratios, maximum likelihood estimation was considered infeasible and instead a pseudo-likelihood estimation procedure that depended only on pairwise parameters was used. We found the Plackett copula did not easily generalise to three dimensions and above, and split larger relative sets into all possible pairs, adjusting the standard errors appropriately. This enabled us to use fully parametric maximum likelihood estimation.

This model also makes clear that for a complex disease, where risk of disease is subject to both genetic and non-genetic factors, λ_R will not be constant, but will depend on an individual's environmental risk factors. It has been implemented in Stata and made available on the internet.[1]

It is disappointing that the model for relative pairs does not extend to trios as was hoped. Such an extension might still be possible were suitable restrictions applied, but it is not obvious what restrictions might be sensible. There was not time in this project to investigate this further, but it is an area that would be interesting to explore.

4.8.2 Estimation of λ_R for leprosy

Pattern across degrees of relationship

In all cases θ is higher for sibling pairs than any other relationship, as discussed recently by Koivisto and Mannila (2001). If θ is measuring purely genetic risk, we would also expect to see θ decline consistently across degree of relationship, as the expected proportion of genes shared decreases from 1/2 to 1/4 and then 1/8. Under condition 1, this does not happen, presumably because of unmeasured non-genetic factors and under condition 2, the estimate of θ for half-sibling pairs clearly does not fit this pattern. This could be due to the considerably higher correlation between non-genetic covariates among half sibling pairs that other second degree pairs, as shown in table 4.4.

Under condition 3, it could be argued that λ_R decreases as we move from sibling pairs to second and third degree pairs, but the estimates for

---

[1] see `http://www-gene.cimr.cam.ac.uk/clayton/software/stata/README.txt`

parents are much lower that for siblings - only just significantly above 1. According to Koivisto and Mannila (2001) such a difference between λ_R values as appears under model 3(a) for siblings and parents is likely to arise because the underlying trait is recessive. They do not examine how differences between second- or third-degree relatives may arise, but figure 4.2 indicates that θ may be higher for half-siblings than aunt/uncles and higher for half aunt/uncles and cousins than grandparents, although these differences are not significant.

The pairwise correlation for all non-genetic risk factors is highest amongst sibling pairs, and the greatest change in estimates between models 1 and 3(a) is for siblings. Estimates are also significantly inflated under model 1 for both parent-offspring and half sibling pairs, and both these pairs also tend to have a high degree of correlation between non-genetic risk factors, particularly household contact.

Magnitude of genetic effect

The results do not provide evidence for a strong genetic susceptibility to leprosy. However, they are consistent with much other evidence that susceptibility to leprosy is under the control of many factors, the strongest of which may be non-genetic, with host genetics playing a small but significant role.

It is possible that the residual risk observed could be accounted for by other, unmeasured, non-genetic factors. These data included the major risk factors that are likely to cluster in families. Exposure was measured using two binary variables: whether an individual shared a household with any MB or PB leprosy case during either of the LEP1 and LEP2 surveys. Given that people tend to change households during their lifetime, these are imperfect measures and in particular do not capture exposure history prior to the KPS. Also, we recognise there was likely to be undetected household contact (Chirwa, 2001), given that our measure of contact was based of just two surveys over ten years. Given the potentially long incubation period of leprosy (measured in years and even decades) and that we are considering cumulative incidence we will certainly have missed earlier household contact, which will be important in older cases. Further, as discussed in section 3.3, the effect of BCG is likely to have been underestimated die to vaccinated individuals being misclassified as unvaccinated. Also, exposure to environmental mycobacteria is also known to affect a person's immune response to infection with *M. leprae*, and this was something we could not account for. It is known that environmental mycobacteria are distributed heterogeneously across Karonga (Chilima, 2001) and tend to cluster in space (and thus within household and family). It is likely that our estimates of λ_R would be further reduced if contact histories and mycobacteria exposure could be accurately accounted for. This may also explain why we do not see the expected pattern of falling λ_R with degree of relationship.

4.9 Summary

Marginal models have previously been applied to disease data to examine the aggregation of disease among relatives. The marginal model proposed in this chapter was applied to present state data from the KPS with the aim of estimating λ_R. We found evidence that $\lambda_R > 1$ for all but grandparent-grandchild pairs (and, specifically, $\lambda_S \in [1.8, 2.0]$), but underestimated effects of marginal covariates and the pattern of λ_R across different relatives indicated this may still be an overestimate.

CHAPTER 5.

AGGREGATION OF DISEASE AMONG RELATIVE TRIOS

5.1 Introduction

Many (100/169) of the pedigrees from Karonga contain more than two affected relatives. The manner in which a genetic disease may aggregate among pairs of relatives has already been widely studied, but less attention has been paid to how disease may aggregate amongst relative trios and above. As described in chapter 4, λ_R is a measure of the excess risk to relatives of cases. Attempts were also made in that chapter to estimate an analogous parameter for relative trios, $\lambda_{R,R}$.

In this chapter, the behaviour of $\lambda_{R,R}$ under one-locus genetic models is explored. This work can also be used to predict the power to detect linkage using affected relative trios, and thus another focus of this chapter is to examine what gain might be had from employing relative trios rather than relative pairs in linkage analysis studies.

The methods used are based on describing the risk of disease given genotype as a function of additive and dominance effects of the alleles and conditioning on the IBD distribution for affected relatives. These *genetic components of variance* and their use for calculating recurrence risk ratios are introduced in section 5.2. Section 5.3 describes restrictions imposed on λ_S by genetic models and section 5.4 describes how $\lambda_{R,R}$ may be expressed in terms of variance components under these models. The power to detect linkage using relative trios is examined in section 5.5 and finally, in section 5.6, the use of relative trios and above in linkage analyses is discussed in relation to the work in this chapter.

5.2 Genetic components of variance and recurrence risks

5.2.1 Components of variance

Assume a quantitative trait is determined by a single locus with n alleles, at population frequencies π_i, $i = 1, \ldots, n$. The mean trait value for an individual with genotype i/j can be written

$$\mu_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij} \tag{5.1}$$

where μ is the population trait mean, α_i and α_j are the additive effects due to alleles i and j respectively, and δ_{ij} is the deviation for genotype i/j. For a binary trait (eg diseased or not), μ_{ij} can be thought of as the risk of disease conditional on genotype. Assume Hardy-Weinberg equilibrium. That is, the population frequency of the unordered genotype i/j is $2\pi_i\pi_j$ for $i \neq j$, or π_i^2 for $i = j$. The α_is are chosen to minimise the sum of squares of the deviations δ_{ij}. This is equivalent to setting

$$\frac{\partial}{\partial \alpha_i} \sum_i \sum_j \pi_i \pi_j \delta_{ij}^2 = \frac{\partial}{\partial \alpha_i} \sum_i \sum_j \pi_i \pi_j (\mu_{ij} - \alpha_i - \alpha_j - \mu)^2 = 0$$

and gives the relations

$$\sum_j \delta_{ij} \pi_j = 0 \tag{5.2}$$

$$\sum_i \alpha_i \pi_i = 0 \tag{5.3}$$

$$\alpha_i = \sum_j \mu_{ij} \pi_j - \mu \tag{5.4}$$

Using these, the genetic variance of the trait can be partitioned into additive and dominance components:

$$\begin{aligned} \sigma^2 &= \sum_i \pi_i \pi_j (\mu_{ij} - \mu)^2 \\ &= \left[2 \sum_i \alpha_i^2 \pi_i \right] + \left[\sum_{i,j} \delta_{ij}^2 \pi_i \pi_j \right] \\ &= \left[\sigma_a^2\right] + \left[\sigma_d^2\right] \end{aligned} \tag{5.5}$$

5.2.2 Recurrence risk ratios

The relative recurrence risk ratio, λ_R, defined in section 2.3.1, is a measure of 'how genetic' a disease is. Let $X_i = 1$ if individual i is affected and 0 otherwise. Then λ_R may also be expressed in terms of the components of genetic variance as

$$\lambda_R = \frac{E(X_1 X_2)}{E(X_1)E(X_2)} = \frac{\mu^2 + \text{cov}(X_1, X_2)}{\mu^2}$$

which, as will be shown in section 5.2.5, gives

$$\lambda_S = \frac{\mu^2 + \sigma_a^2/2 + \sigma_d^2/4}{\mu^2}$$

for siblings.

To consider relative trios, we need to introduce an analogous parameter. The second degree recurrence risk, K_{R_1,R_2}, is defined here as the risk of disease to a person with two affected relatives of type R_1 and R_2,

$$\begin{aligned} K_{R_1,R_2} &= E(X_3 | X_1 = X_2 = 1) \\ &= \frac{E(X_1 X_2 X_3)}{E(X_1 X_2)} \end{aligned}$$

and the second degree recurrence risk ratio, λ_{R_1,R_2}, is defined as the ratio of this to the population risk, ie

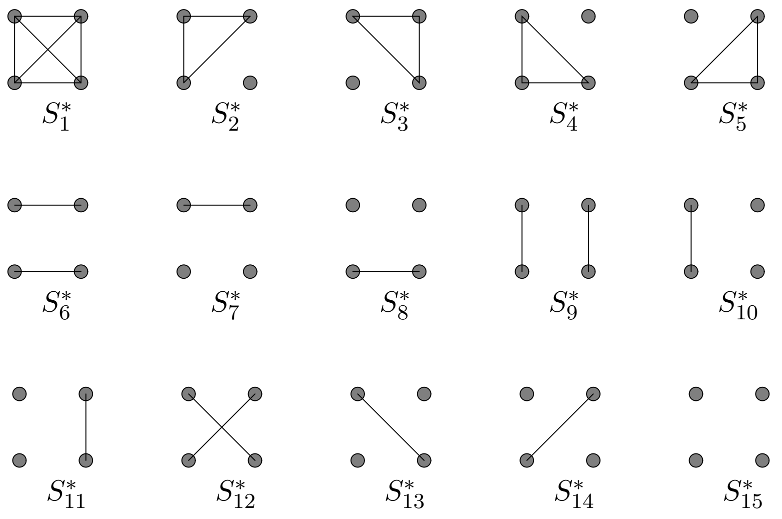
$$\begin{aligned} \lambda_{R_1,R_2} &= \frac{E(X_3 | X_1 X_2)}{E(X_3)} \\ &= \frac{E(X_1 X_2 X_3)}{E(X_1 X_2)E(X_3)}. \end{aligned}$$

λ_{R_1,R_2} is the increase in risk (above the population risk) for an individual who has two affected relatives (of type R_1 and R_2). This parameter may be used to explore the aggregation of disease among relative trios and the relationship between λ_{R_1,R_2} and λ_R may provide further insight still.

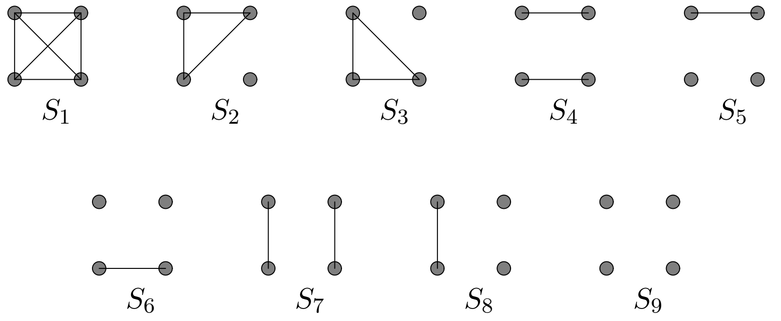
5.2.3 Identity states

The *identity state* of a set of n relatives describes the IBD sharing among the set and can be represented graphically using $2n$ points to represent the maternal and paternal alleles of each relative. Any alleles that are IBD are joined by a solid line.

There are 15 possible identity states for relative pairs, as shown in figure 5.1(a). If we allow maternal and paternal alleles to be rearranged, and individuals swapped, there are 9 *condensed identity states*, shown in figure 5.1(b), three of which contain only non-inbred relatives: S_7, S_8 and S_9. These correspond to a pair sharing two, one or zero alleles IBD respectively.



(a) All possible identity states



(b) Condensed identity states

Figure 5.1: Identity states for relative pairs. Each horizontal pair of vertices represents the maternal and paternal alleles for each relative. Alleles that are IBD are joined by a solid line

An identity state partitions the $2n$ alleles into subsets containing only IBD alleles. Therefore, the number of possible identity states for n relatives is the number of possible partitions of a set with $2n$ members which increases rapidly with n, and for relative trios there are 203 possible IBD sharing configurations (see appendix C for a complete list). If we restrict attention to non-inbred trios and further allow maternal and paternal alleles, or individuals, to be rearranged within the pattern, there are just eight condensed configurations, $T_1, \ldots, T_8$, shown in figure 5.2.

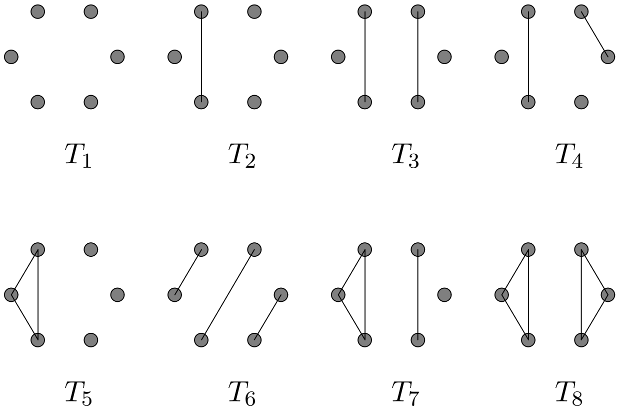


Figure 5.2: Condensed identity states for non-inbred relative trios. Each horizontal pair of vertices represents the maternal and paternal alleles for each relative. Alleles that are IBD are joined by a solid line

There are also many possible relative trios; focus here will be directed to four in particular that arise often in the pedigrees from the KPS: three full siblings (SSS); two full siblings and one half sibling (SSH) three half siblings (HHH) and three cousins (CCC). In the SSH case, two relative risk ratios are possible. $\lambda_{S|S,H}$ will be used to denote the increase in risk for an individual with an affected full sibling and an affected half sibling, while $\lambda_{H|S,S}$ will be used to denote the increase in risk for an individual who is the half sibling of two affected full siblings.

5.2.4 Null identity state distributions

The null distribution for a given relative set (the probability of each identity state given no information about the distribution of any genetic trait among the set) can be found for general pedigrees on application of an algorithm (e.g. Karigl, 1982). However, for non-inbred trios, it is simplest to assign

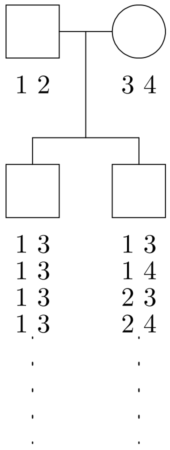


Figure 5.3: Enumerating the null identity state distribution for a sibling pair

unique alleles to each founder in a pedigree and enumerate all possible (and equally likely) IBD configurations among descendents from this - as shown, for example, in figure 5.3 for a sibling pair. The null identity state distributions for selected relative pairs and trios are given in tables 5.1(a) and (b).

5.2.5 Joint probability of disease

For a set of n relatives, let $Y = \prod_i^n X_i$ represent the event that all members of the relative set R are affected. The probability all members of a set are affected conditional on identity state ϕ, $E(Y|\phi)$ may be found using equations (5.1) and (5.2)–(5.5). For relative pairs we find

$$\begin{aligned} E(Y|S_7) &= \sum_{i,j} \pi_i \pi_j \mu_{ij} \mu_{ij} \\ &= \sum_{i,j} \pi_i \pi_j (\mu + \alpha_i + \alpha_j + \delta_{ij})^2 \\ &= \mu^2 + \sigma_a^2 + \sigma_d^2 \end{aligned}$$

Identity State, S_i	$P(S_i)$		
	siblings	2nd degree	3rd degree
S_7	1/4	0	0
S_8	1/2	1/2	1/4
S_9	1/4	1/2	3/4

(a) Relative pairs

Identity state, T_i	$P(T_i)$			
	SSS	SSH	HHH	C
T_1	0	0	0	3/8
T_2	0	3/8	1/2	9/16
T_3	3/16	1/8	0	0
T_4	3/8	1/4	0	0
T_5	0	1/8	1/2	1/16
T_6	0	0	0	0
T_7	3/8	1/8	0	0
T_8	1/16	0	0	0

(b) Relative trios

Table 5.1: Null identity state distribution among selected non-inbred relative sets

and, similarly

$$\begin{aligned} E(Y|S_8) &= \sum_{ijk} \pi_i \pi_j \pi_k \mu_{ij} \mu_{ik} \\ &= \mu^2 + \sigma_a^2/2 \\ E(Y|S_9) &= \sum_{ijkl} \pi_i \pi_j \pi_k \pi_l \mu_{ij} \mu_{kl} \\ &= \mu^2 \end{aligned} \tag{5.6}$$

For relative trios,

$$\begin{aligned} E(Y|T_1) &= \sum_{ijklmn} \pi_i \pi_j \pi_k \pi_l \pi_m \pi_n \mu_{ij} \mu_{kl} \mu_{mn} \\ &= \mu^3 \end{aligned} \tag{5.7}$$

$$
\begin{aligned}
E(Y|T_2) &= \sum_{ijklm} \pi_i\pi_j\pi_k\pi_l\pi_m\mu_{ij}\mu_{ik}\mu_{lm} \\
&= \mu^3 + \mu\sigma_a^2/2 \qquad (5.8) \\
E(Y|T_3) &= \sum_{ijkl} \pi_i\pi_j\pi_k\pi_l\mu_{ij}\mu_{ij}\mu_{kl} \\
&= \mu^3 + \mu(\sigma_a^2 + \sigma_d^2) \qquad (5.9) \\
E(Y|T_4) &= \sum_{ijkl} \pi_i\pi_j\pi_k\pi_l\mu_{ij}\mu_{ik}\mu_{jl} \\
&= \mu^3 + \mu\sigma_a^2 + \sum_{i,j} \alpha_i\alpha_j\delta_{ij}\pi_i\pi_j \qquad (5.10) \\
E(Y|T_5) &= \sum_{ijkl} \pi_i\pi_j\pi_k\pi_l\mu_{ij}\mu_{ik}\mu_{il} \\
&= \mu^3 + 3\mu\sigma_a^2/2 + \sum_i \alpha_i^3\pi_i \qquad (5.11) \\
E(Y|T_6) &= \sum_{ijk} \pi_i\pi_j\pi_k\mu_{ij}\mu_{ik}\mu_{jk} \\
&= \mu^3 + 3\mu\sigma_a^2/2 + 3\sum_{i,j} \alpha_i\alpha_j\delta_{ij}\pi_i\pi_j + \sum_{i,j,k} \delta_{ij}\delta_{ik}\delta_{jk}\pi_i\pi_j\pi_k \qquad (5.12) \\
E(Y|T_7) &= \sum_{ijk} \pi_i\pi_j\pi_k\mu_{ij}\mu_{ij}\mu_{ik} \\
&= \mu^3 + \mu(2\sigma_a^2 + \sigma_d^2) + \sum_i \alpha_i^3\pi_i + \sum_{i,j}(2\alpha_i\alpha_j\delta_{ij} + \alpha_i\delta_{ij}^2)\pi_i\pi_j \\
&\qquad (5.13) \\
E(Y|T_8) &= \sum_{ij} \pi_i\pi_j\mu_{ij}\mu_{ij}\mu_{ij} \\
&= \mu^3 + 3\mu(\sigma_a^2 + \sigma_d^2) + 2\sum_i \alpha_i^3\pi_i + \sum_{i,j}(\delta_{ij}^3 + 6\alpha_i\delta_{ij}^2 + 6\alpha_i\alpha_j\delta_{ij})\pi_i\pi_j \\
&\qquad (5.14)
\end{aligned}
$$

5.2.6 Identity state distribution among affected relatives

Among affected relatives, the distribution of IBD states will differ from the null at loci that influence the risk of disease or are linked to such loci. We can calculate the probability of each state ϕ among an affected relative set R (eg three full siblings) as $k_{R,\phi}p_\phi$ where p_ϕ is the null probability of state ϕ and $k_{R,\phi}$ is a function of the genetic variance of the trait. As before, let

Identity State, S_i	$P(S_i)$		
	sibling pairs	2nd degree pairs	3rd degree pairs
S_7	$\frac{\mu^2}{4\mu^2+2\sigma_a^2+\sigma_d^2}$	0	0
S_8	$\frac{2\mu^2+\sigma_a^2}{4\mu^2+2\sigma_a^2+\sigma_d^2}$	$\frac{2\mu^2+\sigma_a^2}{4\mu^2+\sigma_a^2}$	$\frac{2\mu^2+\sigma_a^2}{8\mu^2+\sigma_a^2}$
S_9	$\frac{\mu^2+\sigma_a^2+\sigma_d^2}{4\mu^2+2\sigma_a^2+\sigma_d^2}$	$\frac{2\mu^2}{4\mu^2+\sigma_a^2}$	$\frac{6\mu^2}{8\mu^2+\sigma_a^2}$

Table 5.2: Identity state distribution for affected sibling, second and third degree relative pairs

X_i represent the disease state of some individual i, with

$$X_i = \begin{cases} 1 & i \text{ affected} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$P(\phi|Y,R) = \frac{E(Y|\phi,R)P(\phi|R)}{E(Y|R)} \tag{5.15}$$

$$= \frac{E(Y|\phi)P(\phi|R)}{\sum_i E(Y|\phi)P(\phi|R)} \tag{5.16}$$

where $P(\phi)$ is the null probability of state ϕ and R the type of relative set.

The identity state distributions for affected sibling, second and third degree relative pairs (found by application of (5.15)) are shown in table 5.2.

Although not listed here, the identity state distribution for any particular relative trio can also be found on application of (5.15).

5.3 Restrictions on λ_S under genetic models

The relationship between genotype and phenotype (eg disease status) may be described by a genetic model. Risch's work and the methods in this chapter compare λ_R or λ_{R_1,R_2} to λ_S, thus treating λ_S as an independent parameter. In fact, λ_S is also dependent on the genetic model and this introduces limitations in the form of valid ranges for λ_S under different models.

Rybicki and Elston (2000) examined the relationship between λ_S and the genotype relative risk, γ, under diallelic models. γ is defined as μ_{11}/μ_{22} where μ_{11} and μ_{22} are the frequency of disease among high and low risk

Model	Frequency of disease-related allele, p				
	0.005	0.01	0.05	0.1	0.5
Dominant	50.56	25.56	5.57	3.25	1.25
Co-dominant	50.75	25.75	5.75	3.25	1.25
Recessive	9338.49	2545.22	110.24	30.35	2.25

Table 5.3: Upper limits for λ_S as $\gamma \to \infty$ under diallelic models. Limits are calculated numerically by setting $\gamma = 10^6$.

homozygotes respectively. They found that under diallelic single and two locus models, even as $\gamma \to \infty$, λ_S was limited according to the frequency of the disease related allele, p. As $p \to 0$, so $\lambda_S \to \infty$, but at moderate values of p, λ_S is limited, and these limits for single locus models (calculated here numerically by setting $\gamma = 10^6$) are shown in table 5.3.

More recently, Schliekelman and Slatkin (2002) examined the relationship between the number of loci in a multiplicative multilocus model and λ_S. They assumed the disease was controlled by L diallelic loci, with all disease-related alleles (at different loci) having equal effect and equal frequency, p. Under this special-case, and with the further restrictions of 'negligible' dominance variance (arbitrarily set as $\lambda_1 \leq 1.11\lambda_S$) and complete penetrance (a person homozygous for the disease-related allele at all susceptibility loci has disease with probability 1), they found λ_S was again restricted by the population prevalence of disease (μ) and the number of loci (L).

As before, the limits on λ_S appear acceptable (eg $\lambda_S < 12$) for a rare disease or rare disease-related alleles, but even moderate values, such as $\lambda_S > 4$ are not possible when the disease is less rare ($\mu = 0.01$) and the number of loci is relatively large ($L > 5$). This is because for such μ, as the number of loci increases, p increases rapidly (eg $p \sim 0.75$ when $L > 5$ and there is no dominance variance), and, as demonstrated in the one- and two-locus case by Rybicki and Elston (2000), λ_S can take only smaller values when p is large.

Schliekelman and Slatkin (2002) used a more restricted model than Risch (1990a,b,c), but showed that λ_S is not an independent parameter as it appears in Risch's work, but is restricted by the underlying disease model.

5.3.1 Two special case models

Later in this chapter, we will consider two specific genetic models:

1. a purely additive model ($\delta_{ij} = 0$)
2. a diallelic model

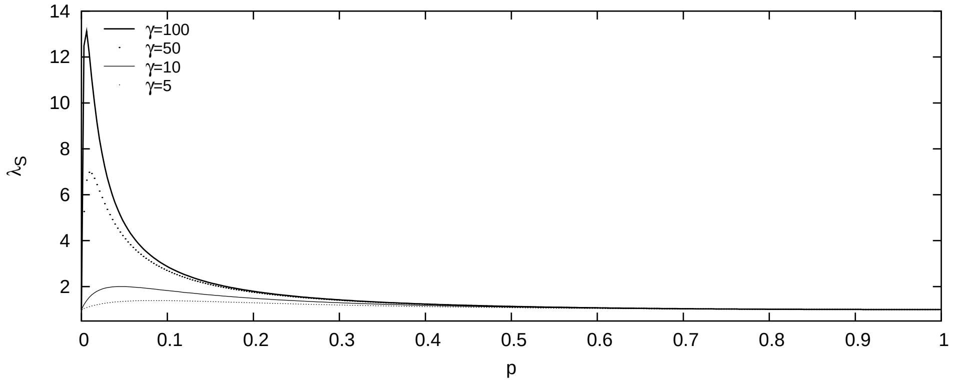
and it is therefore of interest to examine any restrictions on λ_S under these models.

In this discussion, I will be most concerned about bounds that restrict λ_S to moderate values and below (eg $\lambda_S < 10$). This is because we know there will be some restriction on λ_S unless the disease is extremely rare, so a cut-off has to be chosen, above which we think the restriction is reasonable. The choice of what is and what is not a moderate value is fairly arbitrary, but the aim is to focus on cases which are more likely to arise when dealing with complex diseases.

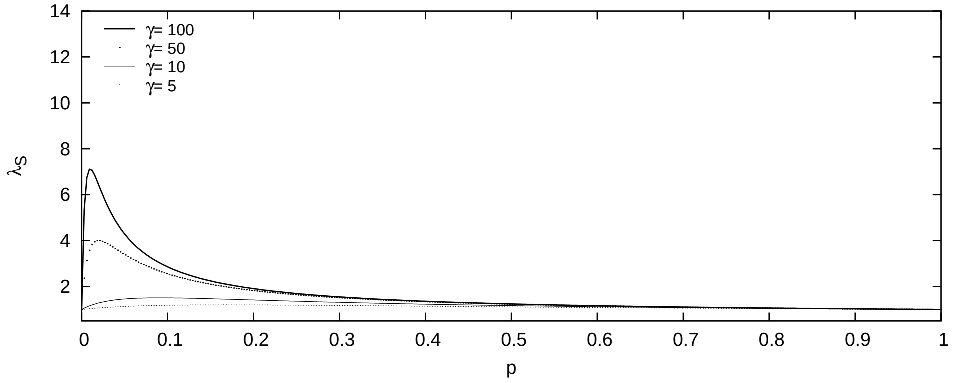
Diallelic models

Bounds for λ_S under diallelic models were given by Rybicki and Elston (2000) and summarised in table 5.3. Note that diallelic models may be specified in terms of three parameters: μ_{11} (penetrance in high-risk homozygotes), μ_{22} (penetrance in low-risk homozygotes) and p. μ_{12} is determined by whether we consider the recessive ($\mu_{12} = \mu_{22}$), co-dominant ($\mu_{12} = (\mu_{11} + \mu_{22})/2$) or dominant ($\mu_{12} = \mu_{11}$) model. Further, we can use the genotype relative risk, $\gamma = \mu_{22}/\mu_{11}$, to summarise the relationship between μ_{11} and μ_{22}.

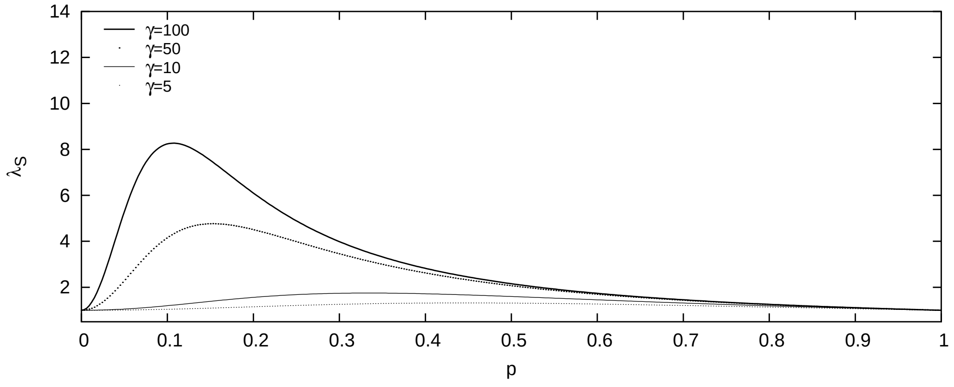
Figure 5.4 shows the one to one relationship between λ_S and p for fixed γ. As γ increases, the range of λ_S increases and the corresponding p value at which λ_S is maximum decreases. As Rybicki and Elston (2000) showed, however, this increase tails off and λ_S is bounded above for any value of p, no matter how large we make γ. Even for quite large γ, λ_S is quite restricted for any diallelic model and any p (eg $\lambda_S < 8$ when $\gamma = 50$). This means that if we are to treat λ_S as an independent parameter like Risch (1990a), we must remember that for larger λ_S, we are implicitly forcing p to be small and γ to be very large indeed.



(a) Dominant model



(b) Codominant model



(c) Recessive model

Figure 5.4: Relationship between λ_S and p under diallelic models for varying genotype relative risk (γ)

No dominance variance assumption

It is harder to fix bounds on this model than on the diallelic model. We introduce another parameter here: the Fisher skewness,

$$\gamma_1 = \frac{2\sum_i \pi_i \alpha_i^3}{\sigma_a^3}.$$

The reasons for this choice of parameter will be given in section 5.4.1; for now, it is sufficient to note that γ_1 is a measure of the (non)centrality of the distribution of μ_{ij}. $\gamma_1 < 0$ and $\gamma_1 > 0$ correspond to negative and positive skew respectively, and $\gamma_1 = 0$ corresponds to symmetry. Although λ_S and γ_1 are related (both are functions of α_i), it is not clear how best to investigate the relationship between them generally.

First, we consider the diallelic case as an example. A codominant diallelic model has no dominance variance and can be parameterised by $\mu_{22} \geq \mu_{11}$ and p, the frequency of (the disease-related) allele 1. Under this model, there is a one-to-one correspondence between p and γ_1:

$$\gamma_1 = \frac{\sqrt{2}(1-2p)}{\sqrt{p(1-p)}}.$$

Let $\mu_{22} = \gamma\mu_{11}$, then setting $\gamma_1 = 0$ fixes $p = 1/2$ so

$$\begin{aligned}\mu &= (\mu_{11} + \mu_{22})/2 = \frac{(1+\gamma)\mu_{11}}{2},\\ \sigma_a^2 &= \frac{(\gamma-1)^2\mu_{11}^2}{8}\end{aligned}$$

and

$$\lambda_S = \frac{\mu + \sigma_a^2/2}{\mu^2} = \frac{5}{4} - \frac{\gamma}{(1+\gamma)^2}.$$

Thus $\lambda_S \geq 1$ increases with γ and tends to 5/4 as $\gamma \to \infty$, as shown in figure 5.5. This bound very low, and although a very special case, indicates that assuming $\gamma_1 = 0$ may not be reasonable. Fixing p at other values fixes γ_1 and bounds for λ_S at these values are shown in table 5.4.

In fact, the bounds under the diallelic case are likely to hold for multiallelic models too. Consider the triallelic case. This can be parameterised by the homozygote means ($\mu_{11} > \mu_{22} > \mu_{33}$) and the frequencies of alleles

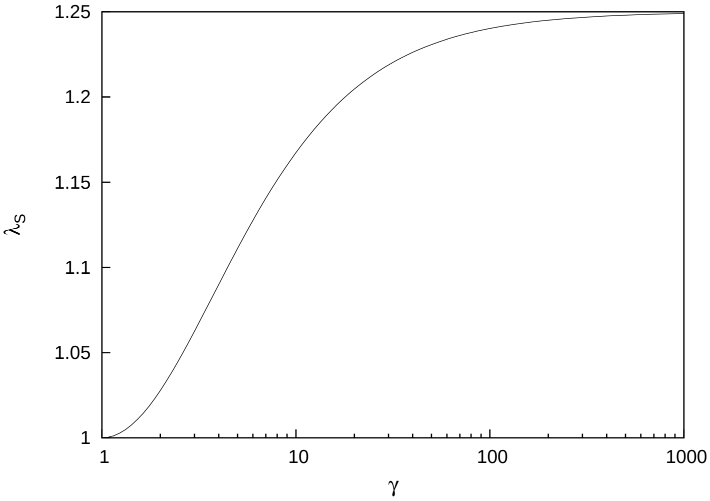


Figure 5.5: λ_S as a function of the genotype relative risk, γ, under a diallelic co-dominant zero skew model

p	γ_1	Upper bound on λ_S
0.001	22.3	250.75
0.005	9.9	50.75
0.010	7.0	25.75
0.029	4.0	9.49
0.050	3.0	5.75
0.092	2.0	3.47
0.100	1.9	3.25
0.211	1.0	1.93
0.500	0.0	1.25

Table 5.4: Bounds on λ_S under a diallelic no dominance variance model

Frequency of allele 1, π_1	Maximum value of γ_1						Upper bound on λ_S
	Frequency of allele 2, π_2						
	0.001	0.01	0.05	0.1	0.2	0.5	
0.001	63.05	26.52	11.54	7.49	4.22	- 0.01	250.74
0.01	26.52	19.39	10.48	7.05	4.02	- 0.11	25.74
0.05	11.54	10.48	7.54	5.54	3.26	- 0.56	5.74
0.1	7.49	7.05	5.54	4.24	2.46	- 1.15	3.24
0.2	4.22	4.02	3.26	2.46	1.15	- 2.46	1.99
0.5	- 0.01	- 0.11	- 0.56	- 1.15	- 2.46	–	1.24

Table 5.5: Maximum values of λ_S and γ_1 for triallelic no dominance variance model

1 and 2 (π_1 and π_2). Again, the no dominance variance assumption implies $\mu_{ij} = (\mu_{ii} + \mu_{jj})/2$. Graphical examination of γ_1 and λ_S showed that the bounds λ_S and γ_1 were again dependent on the frequency of the highest risk allele, π_1. The frequency and size of effect of genotype 2/2 did not change the bound itself, but varied how quickly that bound was reached (in terms of increasing μ_{11}/μ_{33}). The maximum values of γ_1 and λ_S for different π_1 and π_2 are shown in table 5.5. As $\pi_2 \rightarrow 0$, λ_S increases and the model tends to the diallelic case above. The maximum value of γ_1 tends to its value under the diallelic case for $\pi_1 = p$, confirming this. Of greatest interest is that the upper bound on λ_S is the same as in the diallelic case (when $\pi_1 = p$). Although not a mathematical proof, this does indicate that the bounds for λ_S under the diallelic case will hold for multiallelic no dominance variance models too.

So, as before, bounds on λ_S are tighter when p is smaller. It appears that very small values of γ_1 are unrealistic for disease genes which we would hope to detect with genome screens of moderate size, ie diseases with $\lambda_S \geq 1.5$ at minimum. As the number of alleles increases, it is likely that the range of possible values for γ_1 increases as it does when we move from the diallelic to triallelic cases. However, note that γ_1 can be much larger than was shown in the examples of skewed distributions in figure 5.6 when disease related the allele(s) are rare.

Overall, these examples help explain the relationship between the upper bound for λ_S and other genetic parameters. They will be useful when interpreting the results described in the following sections.

5.4 Calculation of recurrence risks under two special-case models of inheritance

The genetic components of variance for relative pairs can be expressed using only three unknown quantities - μ, σ_a^2 and σ_d^2. For a relative trio we have a further five: $\sum_i \alpha_i^3 \pi_i$; $\sum_{i,j} \alpha_i \alpha_j \delta_{ij} \pi_i \pi_j$; $\sum_{i,j} \alpha_i \delta_{ij}^2 \pi_i \pi_j$; $\sum_{i,j} \delta_{ij}^3 \pi_i \pi_j$ and $\sum_{i,j,k} \delta_{ij} \delta_{ik} \delta_{jk} \pi_i \pi_j \pi_k$. These will not generally simplify. Here, two special cases are considered:

1. a no dominance variance model ($\delta_{ij} = 0$)
2. a diallelic model

5.4.1 No dominance variance model

Set $\delta_{ij} = 0$, then $\sigma_d^2 = 0$ and the expressions in equations (5.7)– (5.14) involve just three unknown quantities - μ, σ_a^2 and $\sum_i \alpha_i^3 \pi_i$. This is the no dominance variance (NDV) assumption and implies that the effect for any genotype can be expressed as the sum of the effects of its alleles.

In the two-dimensional case, reparamaterisation using the relation

$$\sigma_a^2 = 2\mu^2(\lambda_S - 1)$$

allows the unknown additive variance, σ_a^2, to be expressed as a function of a more familiar parameter, λ_S. Similarly, in the three dimensional case, we want to express $\sum_i \pi_i \alpha_i^3$ in terms of some more familiar parameter, and some measure of skewness seems most natural.

Measures of skewness

In a non-symmetric distribution, the mean and mode are not equal. One measure of skewness is the scaled difference between the median and the mode,

$$\frac{\text{mean} - \text{mode}}{\sigma},$$

but the mode is not generally simple to find analytically. Other measures include

Fisher skewness $$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Pearson skewness $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \gamma_1^2$$

Pearson skewness coefficient $$S_K = \frac{[\text{mean}] - [\text{median}]}{\sigma}$$

Bowley skewness $$S_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}$$

where μ_n denotes the n^{th} central moment and Q_n the n^{th} quartile. These measures are shown for data simulated from four increasingly skewed distributions in figure 5.6.

It was decided to use γ_1 here, because it is a linear function of the third moment, μ_3, as is the unknown $\sum_i \pi_i \alpha_i^3$. The third central moment of the trait is given by

$$\sum_{ij} \pi_i \pi_j (\mu_{ij} - \mu)^3 = 2 \sum_i \pi_i \alpha_i^3$$

so the Fisher skewness is

$$\gamma_1 = \frac{2 \sum_i \pi_i \alpha_i^3}{\sigma_a^3}.$$

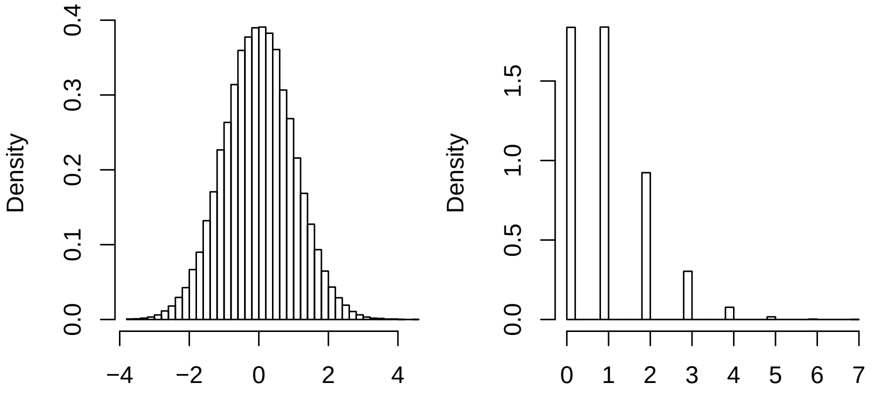
All unknowns can now be expressed in terms of μ, σ_a^2 and γ_1 and we can parameterise using γ_1 and λ_S, giving, for example,

$$\lambda_{S,S} = \frac{-4 + 6\lambda_S + \gamma_1(\lambda_S - 1)\sqrt{2(\lambda_S - 1)}}{2\lambda_S}$$

μ does not appear in the above expression because it has cancelled, and this also applies for all λ_{R_1,R_2} considered. We are thus left with just two free parameters: λ_S and γ_1.

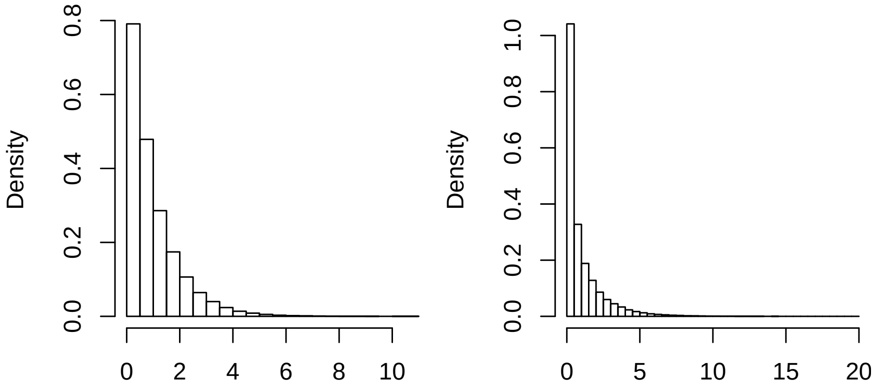
5.4.2 Diallelic model

A trait described by a diallelic model can have three distinct genetic types, and so has four independent parameters: the frequency of the disease-related



(a) $N(0,1)$:
$\gamma_1 = 0$, $\beta_1 = 0$, $S_k = 0$, $S_b = 0$

(b) $P(1)$:
$\gamma_1 = 1$, $\beta_1 = 1$, $S_k = 0$, $S_b = 0$



(c) $\Gamma(1,1)$:
$\gamma_1 = 2$, $\beta_1 = 4$, $S_k = 0.3$, $S_b = 0.3$

(d) $\chi^2(1)$:
$\gamma_1 = 2.8$, $\beta_1 = 8$, $S_k = 0.4$, $S_b = 0.4$

Figure 5.6: Examples of different measures of skewness for simulated data from four distributions

allele, p, and three trait means

$$\begin{aligned}\mu_{11} &= \mu + 2\alpha_1 + \delta_{11}\\ \mu_{12} &= \mu + \alpha_1 + \alpha_2 + \delta_{12}\\ \mu_{22} &= \mu + 2\alpha_2 + \delta_{22}.\end{aligned}$$

These equations may be solved subject to the restrictions (5.2)– (5.4) to give

$$\begin{aligned}\mu &= p^2\mu_{11} + 2p(1-p)\mu_{12} + (1-p)^2\mu_{22}\\ \alpha_1 &= (1-p)(p\mu_{11} + (1-2p)\mu_{12} - (1-p)\mu_{22})\\ \alpha_2 &= p(-p\mu_{11} - (1-2p)\mu_{12} + (1-p)\mu_{22})\\ \delta_{11} &= (1-p)^2(\mu_{11} - 2\mu_{12} + \mu_{22})\\ \delta_{12} &= -p(1-p)(\mu_{11} - 2\mu_{12} + \mu_{22})\\ \delta_{22} &= p^2(\mu_{11} - 2\mu_{12} + \mu_{22})\end{aligned}$$

which allows all the unknown quantities listed at the start of this section to be expressed in terms of μ_{11}, μ_{12}, μ_{22} and p. We may reduce the number of parameters further by considering three specific diallelic models which fix μ_{12} in terms of μ_{11} and μ_{22}: dominant ($\mu_{12} = \mu_{11}$), co-dominant ($\mu_{12} = (\mu_{11} + \mu_{22})/2$) and recessive ($\mu_{12} = \mu_{22}$). We then express μ_{11} and μ_{22} in terms of λ_S and μ so that λ_{R_1,R_2} is then a function of μ, λ_S and p, giving, for example,

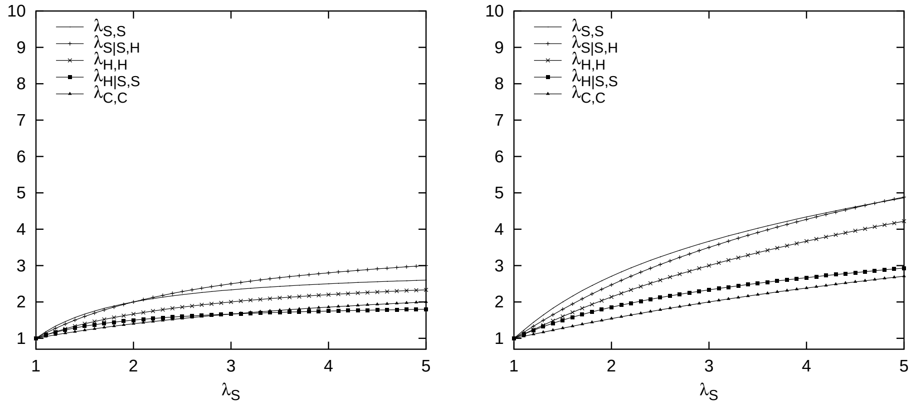
$$\lambda_{S,S} = \frac{(2p-1)(\lambda_S - 1)\sqrt{(\lambda_S - 1)p(1-p)} + 2(3\lambda_S - 2)p(1-p)}{2p(1-p)\lambda_S}$$

under a codominant model. Again, μ has cancelled and we are left with two free parameters: λ_S and p.

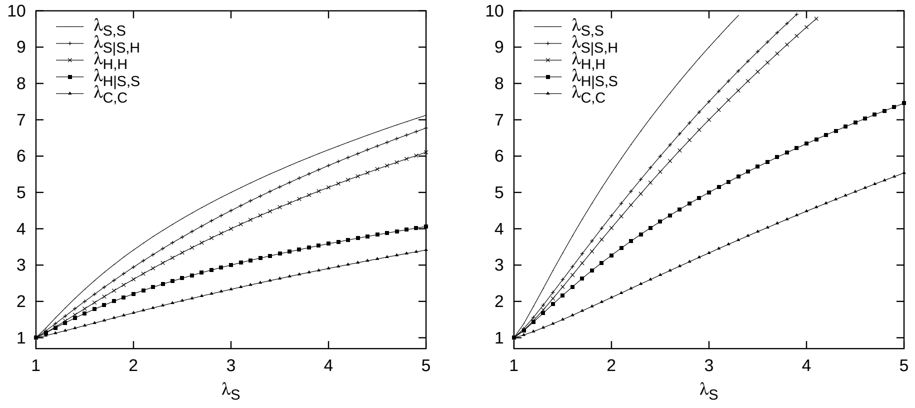
5.4.3 Results

Plots of the second degree $\lambda_{R,R}$ are shown for four relative trios in figure 5.7 under the NDV model and in figure 5.8 under diallelic models.

As might be expected, under both the NDV and diallelic models, risk of disease in any person with two affected relatives increases with λ_S. Note that if we had not investigated the bounds on λ_S, we would find $\lambda_{S,S}$ to actually be *less than* λ_S for $\lambda_S > 2.5$ and an unskewed distribution ($\gamma_1 = 0$)

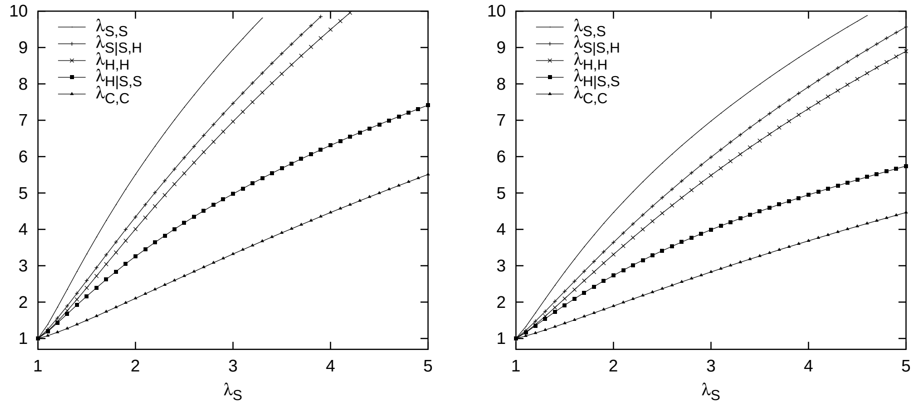


(a) $\gamma_1 = 0$ (b) $\gamma_1 = 2$

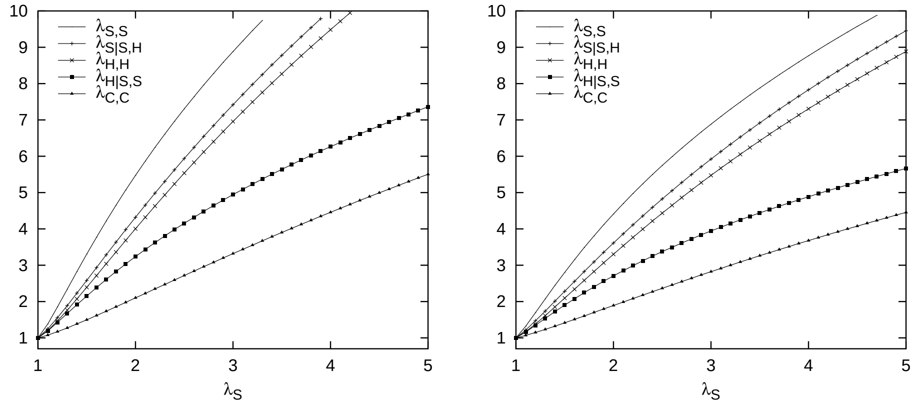


(c) $\gamma_1 = 4$ (d) $\gamma_1 = 10$

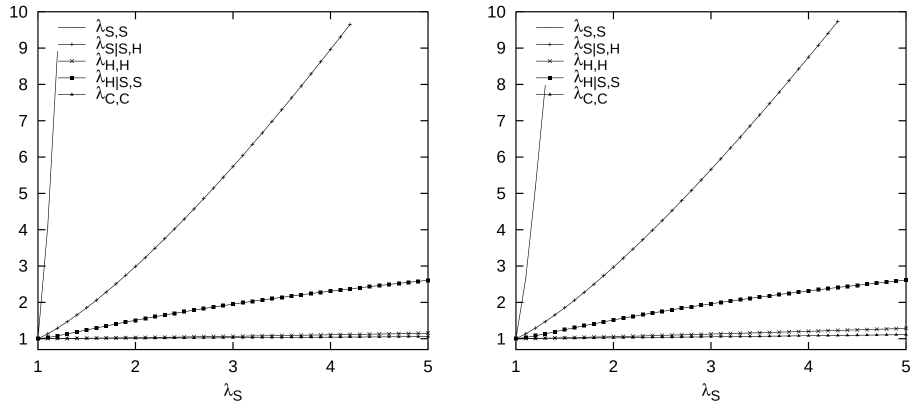
Figure 5.7: Second degree $\lambda_{R,R}$s under the no dominance variance model



(a) dominant; $p = 0.005$ (b) dominant; $p = 0.01$



(c) co-dominant; $p = 0.005$ (d) co-dominant; $p = 0.01$



(e) recessive; $p = 0.005$ (f) recessive; $p = 0.01$

Figure 5.8: Second degree $\lambda_{R,R}$s under diallelic models

under the NDV model. This would be very surprising if true, because we would expect a person's risk of disease given they have two affected siblings to be at least as great as if they had one affected sibling.

In fact, we showed in section 5.3.1 that $\lambda_S \leq 5/4$ for a zero skew diallelic model and that this limit is likely to also hold for multiallelic models. Assuming it does, it would be impossible that $\lambda_S > 2.5$ when $\gamma_1 = 0$, indicating that this apparently unexpected result is due to λ_S exceeding its upper bound. Similarly for $\gamma_1 = 2$, $\lambda_S < 3.5$.

Generally, under the NDV model, λ_{R_1,R_2} increases more rapidly as γ_1 increases, indicating that the more skewed the distribution of the genetic trait (or the greater the difference in risk between genetically susceptible and non-susceptible individuals), the more disease is likely to cluster in families.

Although λ_{R_1,R_2} increases more rapidly under the dominant model than under the co-dominant model, there is only a very small difference for fixed p. This is because the disease-related allele has been assumed to be rare ($p = 0.01$ or $p = 0.005$) so the homozygote will be extremely rare (frequency p^2). This means the model is dominated by the relationship between the heterozygote (frequency $p(1-p)$) and homozygote disease-resistant genotype (frequency $(1-p)^2$). Under the dominant and codominant diallelic models and the NDV model, the following order holds for valid ranges of λ_S:

$$\lambda_{S,S} > \lambda_{S|S,H} > \lambda_{H,H} > \lambda_{H|S,S} > \lambda_{C,C}$$

Under the recessive model however, we find $\lambda_{S,S}$ increases extremely rapidly with respect to λ_S (for example, when $p = 0.005$, $\lambda_{S,S} = 9$ while $\lambda_S < 1.5$). $\lambda_{S|S,H}$ increases at a similar rate to that under the dominant models over the range of λ_S considered here, but the curve is now concave, as oppose to convex, so this increase will be more dramatic for large λ_S. The curves for $\lambda_{H,H}$ and $\lambda_{C,C}$ are almost flat since these trios share only one ancestor which makes it unlikely all three will inherit two copies of a rare disease allele. $\lambda_{H|S,S}$ is slightly higher, but only reaches 2.5 at $\lambda_S = 5$.

Risch (1990a,b,c) showed that λ_R could be expressed simply as a function of the mean and variance of the genetic trait. This work has shown that $\lambda_{R,R}$ (and so the aggregation of disease among relative trios) depends on further parameters. While this makes prediction of how disease may aggregate more complicated with simplified models, it does imply that relative trios (and above) may offer more power to discriminate between different genetic models than relative pairs.

5.5 Power to detect linkage

5.5.1 Using relative pairs

Risch (1987) showed that the probability an affected relative pair of type R will share i alleles IBD at a trait locus, z_{Ri}, can be written in terms of λ_R values and is given by

$$\begin{aligned} z_{R0} &= \alpha_{R0}/\lambda_R \\ z_{R1} &= \alpha_{R1}\lambda_O/\lambda_R \\ z_{R2} &= \alpha_{R2}\lambda_M/\lambda_R \end{aligned}$$

where α_{Ri} is the null probability of sharing i alleles IBD.

Risch (1990b) later examined the power of affected relative pairs in non-parametric linkage analysis using a maximum lod score statistic

$$T = n_0 \log_{10}\left(\frac{n_0}{N\alpha_{R0}}\right) + (N - n_0)\log_{10}\left(\frac{N - n_0}{N - N\alpha_{R0}}\right)$$

where n_0 is the number of N affected relative pairs sharing 0 alleles IBD. The power to detect linkage using T can be shown to be a function of λ_R and θ (the recombination fraction between marker and disease loci). Under a single gene or additive multilocus model,

$$\lambda_1 - 1 = 2(\lambda_2 - 1) = 4(\lambda_3 - 1).$$

If, further, there is NDV, $\lambda_S = \lambda_1$ and the power to detect linkage is a function of λ_1 and θ alone. Generally, higher degree relative pairs provide more power to detect linkage than lower degree pairs, but bigger differences between different relative pairs are found when $\theta > 0$. In this case, distant relatives are not automatically preferred to close ones because there are more opportunities for crossing-over between marker and trait locus. Distinctions also appear between different relative pairs of the same degree. For example, when $\theta = 0.05$, grandparent-grandchild pairs have greatest power when

$\lambda_S(=\lambda_O) < 4$ and first cousins are preferred for larger λ_S, while when $\theta = 0.1$, grandparent-grandchild pairs are always more powerful than first cousins.

5.5.2 Using relative trios

A similar approach can be applied to relative trios. Suppose N relative sets of type R are observed, n_i of whom are in state ϕ_i, $i = 1, \ldots, N_R$. Under the null, the n_i follow a multinomial distribution with probabilities $P(\phi_i|R)$. Like Risch, we can use a likelihood ratio to test for departure from the null distribution.

$$\Lambda = \frac{\prod_i (n_i/N)^n}{\prod_i P(\phi_i|R)^n}$$

is a likelihood ratio and $\log_{10}\Lambda$ is a lod score, so a criterion of the form $\log_{10}\Lambda > K$ can be used to test for linkage.

The simplest case to consider is half siblings, as there are only two possible identity states for half sibling pairs or trios.

No recombination between marker and trait loci

Initially, set $\theta = 0$ to simplify matters. For affected half sibling (HS) pairs, we have (from (5.15)– (5.6))

$$P(S_8) = p_s = \frac{2\mu^2 + \sigma_a^2}{4\mu^2 + \sigma_a^2}$$
$$P(S_9) = 1 - p_s = \frac{2\mu^2}{4\mu^2 + \sigma_a^2}$$

and for affected HS trios, we have (from (5.15), (5.7)–(5.14) and table 5.1(b))

$$P(T_5) = p_t = \frac{\mu^3 + 3\mu\sigma_a^2/2 + \sum \alpha_i^3\pi}{4\mu^3 + 3\mu\sigma_a^2 + \sum \alpha_i^3\pi}$$
$$P(T_2) = 1 - p_t = \frac{3\mu^3 + 3\mu\sigma_a^2/2}{4\mu^3 + 3\mu\sigma_a^2 + \sum \alpha_i^3\pi}.$$

Suppose we observe N affected HS pairs or trios, n of whom are in identity state ϕ_1 (and $N - n$ in state ϕ_2). Under the null, n has a binomial distribution

$$n \sim B\left(N, P(\phi_1)\right).$$

The power to detect linkage can be determined by finding the number n' such that when $n \geq n'$, $\log_{10} \Lambda > K$ and calculating the probability that a realization from a $B(N, P(\phi_1|Y))$ distribution exceeds n'. This will be a function of μ, σ_a^2 and $\sum_i \alpha_i^3 \pi_i$.

Allowing for recombination between marker and trait loci

The above work assumes that we can measure IBD sharing at the trait locus, either because there is no recombination between the trait and marker loci or because we are typing the trait locus directly. This is obviously unrealistic. Risch showed how his method could be extended to allow for recombination between marker and trait. We use the same strategy here, described below in terms of identity state distributions.

Let $\phi_1^{(t)}$ and $\phi_1^{(m)}$ denote the events that the HS pairs or trios are in identity state T_5 or S_8 at the trait and marker locus, respectively, and let θ denote the recombination fraction between the two. We showed above that the power to detect linkage was the probability that an observation from a $B(N, P(\phi_1^{(t)}|Y))$ distribution exceeded n', where n' was chosen such that $\log +10\Lambda > 3$ for $n \geq n'$. The same method applies if we allow for recombination, but, since we observe marker and not trait genotypes, we must use $P(\phi_1^{(m)}|Y)$ in place of $P(\phi_1^{(t)}|y)$. We can use Bayes theorem to write

$$\begin{aligned} P(\phi_1^{(m)}|Y) &= \frac{P(Y|\phi_1^{(m)})P(\phi_1^{(m)})}{P(Y)} \\ &= \frac{P(\phi_1^{(m)})}{P(Y)} \sum_{i=1}^{2} P(Y|\phi_i^{(t)})P(\phi_i^{(t)}|\phi_1^{(m)}). \end{aligned}$$

The only unknown in the above expression is $P(\phi_i^{(t)}|\phi_1^{(m)})$. This can be found by considering the number of recombinations between the loci for $\phi_1^{(m)}$ that would lead to state $\phi_i^{(t)}$, and the probabilities are given in table 5.6.

As above, we reparamaterise the problem using λ_S and γ_1. However, this introduces another parameter σ_d^2. This can be dealt with either by setting $\sigma_d^2 = 0$ (ie applying the NDV assumption again) or by considering diallelic models. Under either assumption, μ cancels, and the power will be a function of λ_S and γ_1.

	Identity state at marker locus	
	$S_8^{(m)}$	$S_9^{(m)}$
$P(S_8^{(t)})$	$\theta^2 + (1-\theta)^2$	$2\theta(1-\theta)$
$P(S_9^{(t)})$	$2\theta(1-\theta)$	$\theta^2 + (1-\theta)^2$

(a) HS pairs

	Identity state at marker locus	
	$T_2^{(m)}$	$T_5^{(m)}$
$P(T_2^{(t)})$	$1-\theta+\theta^2$	$3\theta(1-\theta)$
$P(T_5^{(t)})$	$\theta(1-\theta)$	$1-3\theta(1-\theta)$

(b) HS trios

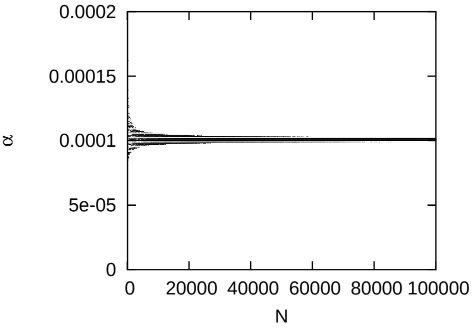
Table 5.6: Probability of identity state at the trait locus conditional on identity state at the marker locus

5.5.3 Results

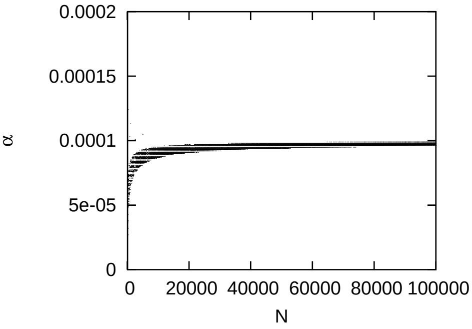
It was decided to examine power to detect linkage using the criterion $\log_{10} \Lambda > 3$ and 60, 120 or 180 pairs of half siblings (HS). To keep the number of people required for genotyping comparable, power was examined for 40, 80 or 120 HS trios. For these numbers of HS pairs and trios, the threshold for declaring significant linkage is

$$n' = \begin{cases} 45 & N = 60 \\ 81 & N = 120 \\ 115 & N = 180 \end{cases} \text{(pairs)} \qquad \text{or} \qquad \begin{cases} 22 & N = 40 \\ 36 & N = 80 \\ 49 & N = 120 \end{cases} \text{(trios)}$$

Asymptotic significance levels are not easy to calculate algebraically given the form of Λ. They have been calculated numerically (using a normal approximation to the binomial distribution) for $N \leq 1,000,000$ and tend towards 10^{-4} for both pairs and trios (see figure 5.9). However, for smaller N, significance levels may vary from this asymptotic value and so exact significance levels for the above thresholds have been calculated and found to



(a) HS pairs



(b) HS trios

Figure 5.9: Asymptotic significance levels for $\log_{10} \Lambda = 3$ *threshold*

be

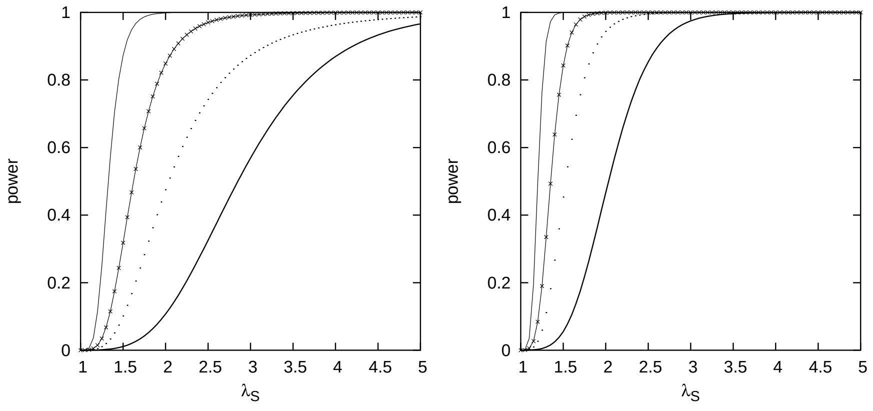
$$\alpha = \begin{cases} 6.7 \times 10^{-5} & N = 60 \\ 7.9 \times 10^{-5} & N = 120 \\ 1.2 \times 10^{-4} & N = 180 \end{cases} \text{(pairs)} \quad \text{or} \quad \begin{cases} 5.9 \times 10^{-5} & N = 40 \\ 7.9 \times 10^{-5} & N = 80 \\ 1.1 \times 10^{-4} & N = 120 \end{cases} \text{(trios)}$$

No recombination between marker and trait loci

Under the idealised situation of no recombination between marker and trait loci, the power to detect linkage using relative pairs is a function of λ_S alone, and a function of λ_S and γ_1 for trios. Power calculated under the NDV model is presented for $\gamma_1 = 0, 2, 10$ and λ_S varying from 1 to 5 in figure 5.10. The results for dominant and co-dominant diallelic models are shown in figure 5.11. The results for the recessive model are not shown as the power to detect linkage under this model was, as might be expected, extremely low, since half siblings cannot share two alleles IBD.

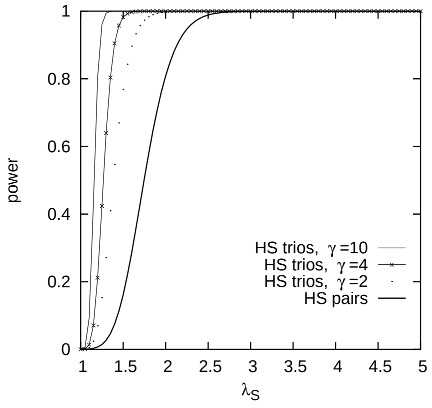
Half sibling (HS) trios generally provide more power to detect linkage than pairs, and the increase in power can be substantial. For example, with only 60 HS pairs, we have 80% power to detect a gene responsible for a λ_S of nearly 4. But genotyping the same number of individuals (40 HS trios) provides 80% power to detect a gene responsible for a λ_S of under 2 under a model with NDV and some positive skew ($\gamma_1 = 2$) or even a gene responsible for a λ_S of under 1.5 under a diallelic model with a moderately rare disease-related allele ($p < 0.05$).

As before, results for the diallelic dominant and co-dominant models are very similar. The increase in power available by using HS trios under these models is substantially higher than under the NDV model, particularly when the disease-related allele is rare. However, the difference in power falls as the disease-related allele becomes more common. This is because as the allele becomes more common, so the chance that affected half siblings inherit it independently increases. Theoretically, HS trios actually provide *less* power when p and λ_S are both large enough (eg $p > 0.7$ and $\lambda_S > 2$), but as shown earlier, there is an upper limit on λ_S and when $p = 0.7$, this limit is $\lambda_S < 1.04$ under the dominant model and $\lambda_S < 1.1$ under the co-dominant model. This means that under all parameter values which are valid for diallelic single-locus models, HS trios are more powerful than HS pairs in a linkage study.



(a) 60 half sibling pairs or 40 half sibling trios

(b) 120 half sibling pairs or 80 half sibling trios



(c) 180 half sibling pairs or 120 half sibling trios

Figure 5.10: Power to detect linkage using half sibling pairs and trios: no dominance variance model

Allowing for recombination between marker and trait loci

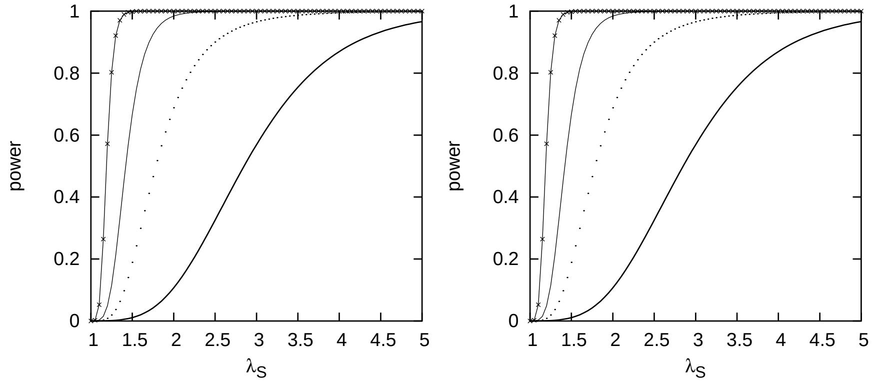
Consider recombination fractions $\theta = 0$, 0.05, 0.1. Currently, genome screens are generally conducted with a marker spacing of at most 10cM and so we would expect the trait locus to be within a recombination fraction of 0.05 of some marker. To save space, only selected results are presented here. The results are presented for 40 trios/60 pairs and 120 pairs/180 trios for $\gamma = 2$ and $\gamma = 10$ under the NDV model (figures 5.12 and 5.13) and for $p = 0.001, 0.05$ under the diallelic codominant model (figures 5.14 and 5.15). (Results under the diallelic dominant model are almost identical).

Although power decreases when we we allow recombination between the trait and marker locus, trios remain more powerful than pairs. This decrease in power is reasonably small when the disease-related allele is rare (in the diallelic case) or the genetic distribution has strong skew ($\gamma = 10$ under the NDV distribution), but can be quite substantial when the allele is more common or the distribution less skewed. This decrease in power also has less effect when more pairs or trios are included in the analysis.

A useful alternative when comparing power is to compare not how much power is available for N pairs/trios, but how many pairs/trios are required to reach a given level of power. The power to detect linkage for increasing N (the number of pairs/trios) and fixed λ_S under NDV and codominant models is shown in figure 5.16 and the number of pairs/trios required to reach 50%, 80% and 90% power are given in table 5.7. There are fewer affected trios required to reach the same levels of power as pairs, particularly when λ_S is small. For example, only 37 trios are required to reach 80% power under an NDV model with strong skew ($\gamma_1 = 10$) compared to over 500 pairs when $\lambda_S = 1.5$. When $\lambda_S = 5$, however, only 15 trios are needed compared with 44. Again, these numbers are calculated under the idealised situation of no recombination between marker and trait and complete IBD sharing information. Still, they illustrate the dramatic increase in power available when λ_S is small.

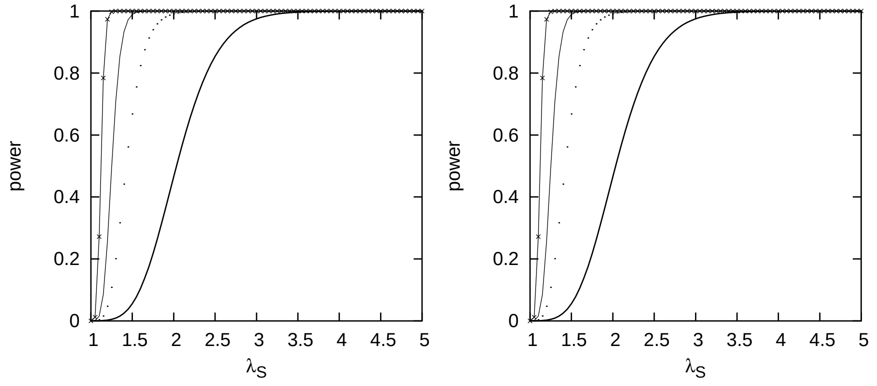
5.6 Discussion

Patterns of disease aggregation among unaffected relative pairs have long been studied in order to answer questions such as ‘what is the probability one individual is of a certain type given the type of another related individ-



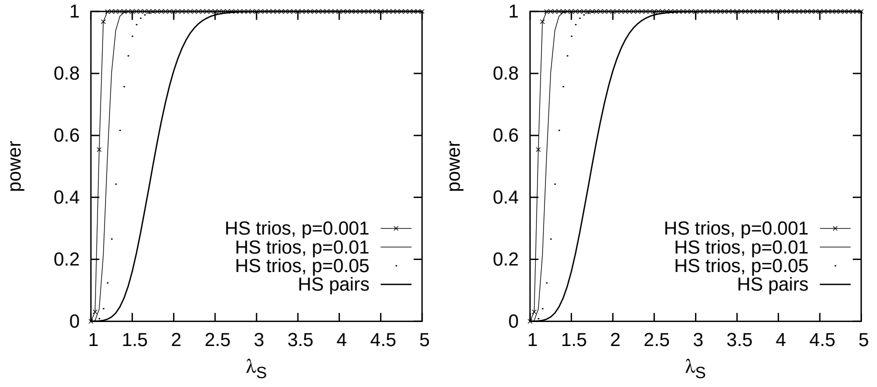
(a) dominant model: 60 pairs or 40 trios

(b) co-dominant model: 60 pairs or 40 trios



(c) dominant model: 120 pairs or 80 trios

(d) co-dominant model: 120 pairs or 80 trios



(e) dominant model: 180 pairs or 120 trios

(f) co-dominant model: 180 pairs or 120 trios

Figure 5.11: Power to detect linkage using half sibling pairs and trios: diallelic dominant and co-dominant models

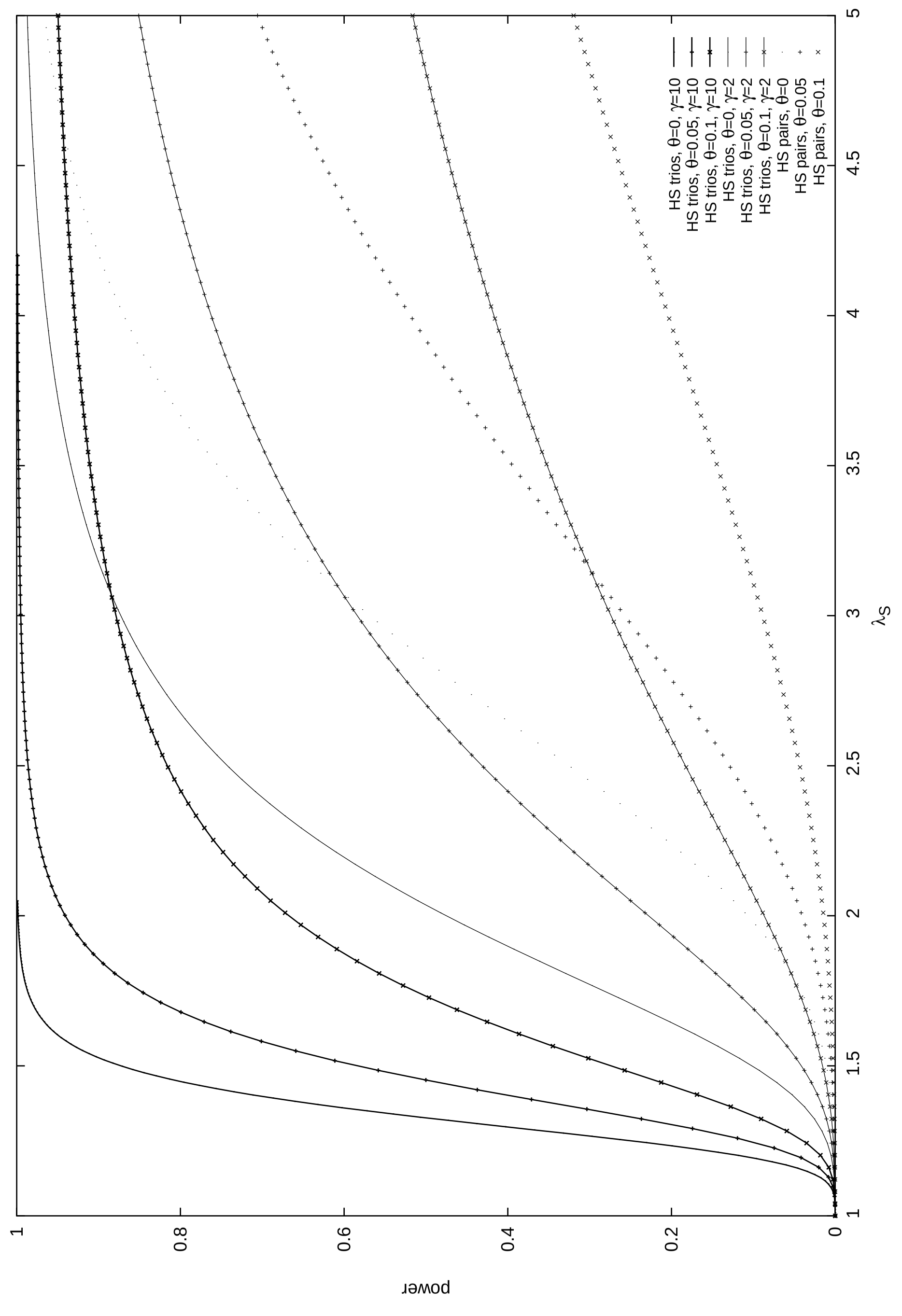


Figure 5.12: Power to detect linkage: no dominance variance model, $\theta \geq 0$ I. 60 HS pairs or 40 HS trios

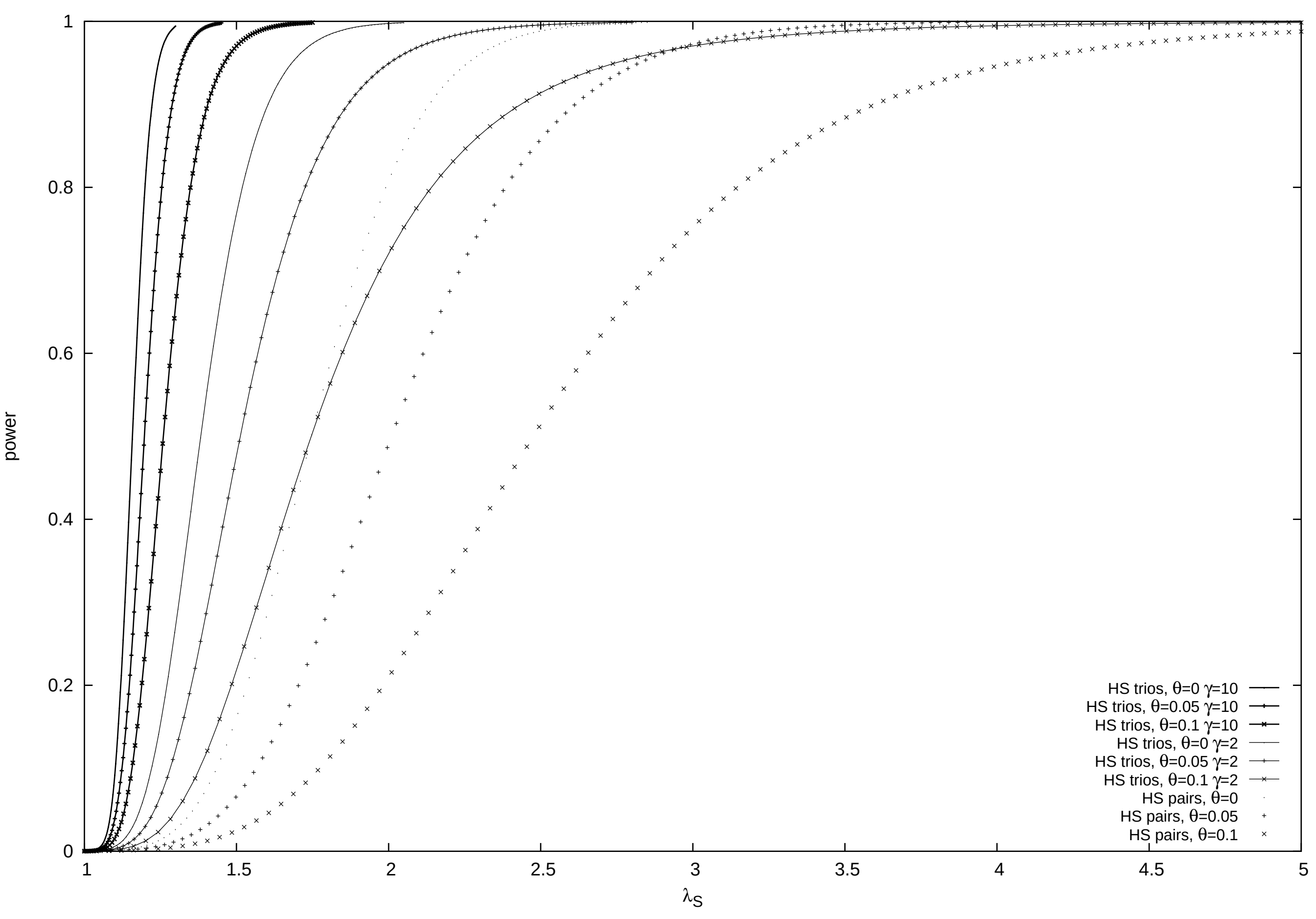


Figure 5.13: Power to detect linkage: no dominance variance model, $\theta \geq 0$ II. 120 HS pairs or 180 HS trios

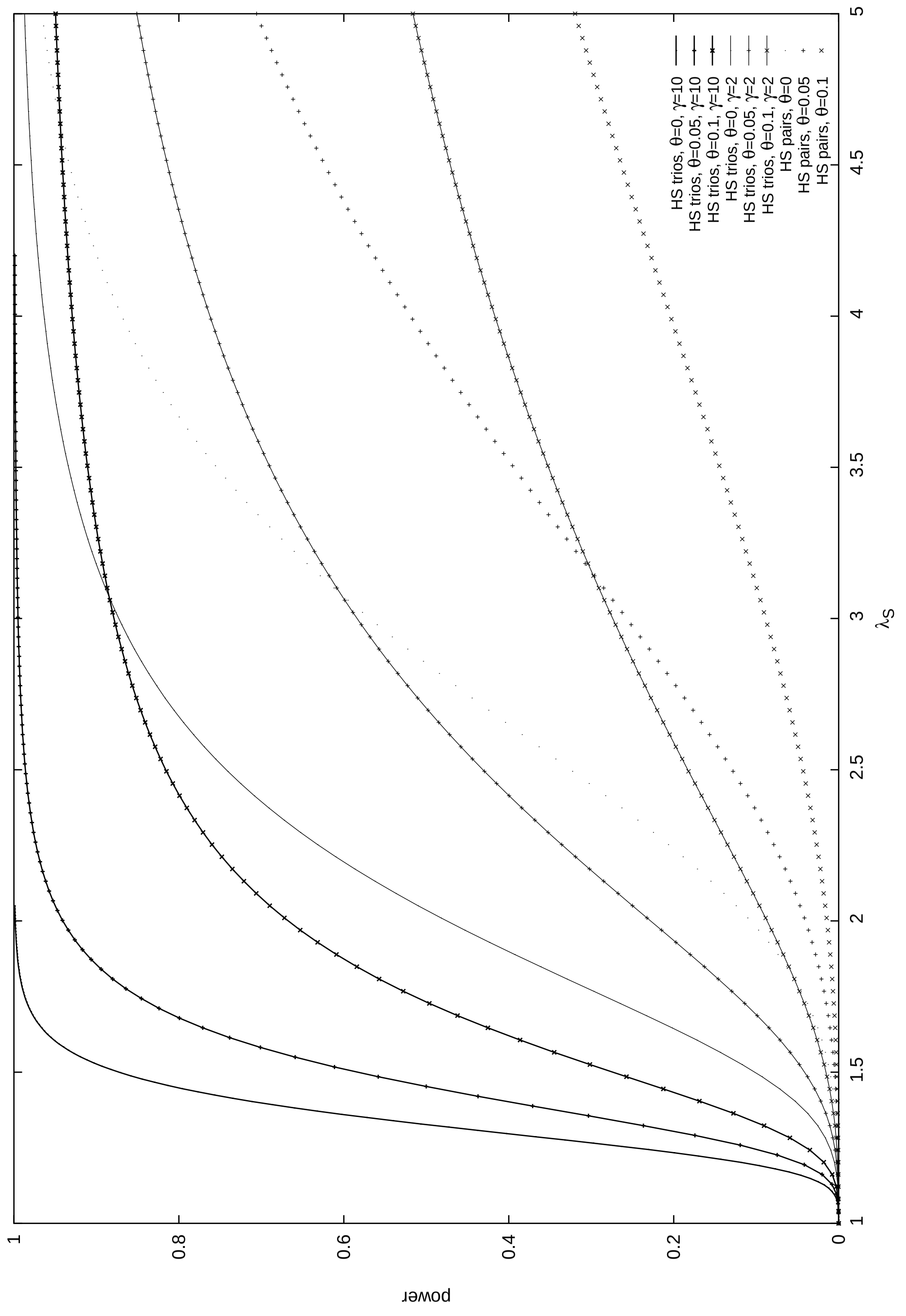


Figure 5.14: Power to detect linkage: diallelic codominant model, $\theta \geq 0$ I. 60 HS pairs or 40 HS trios

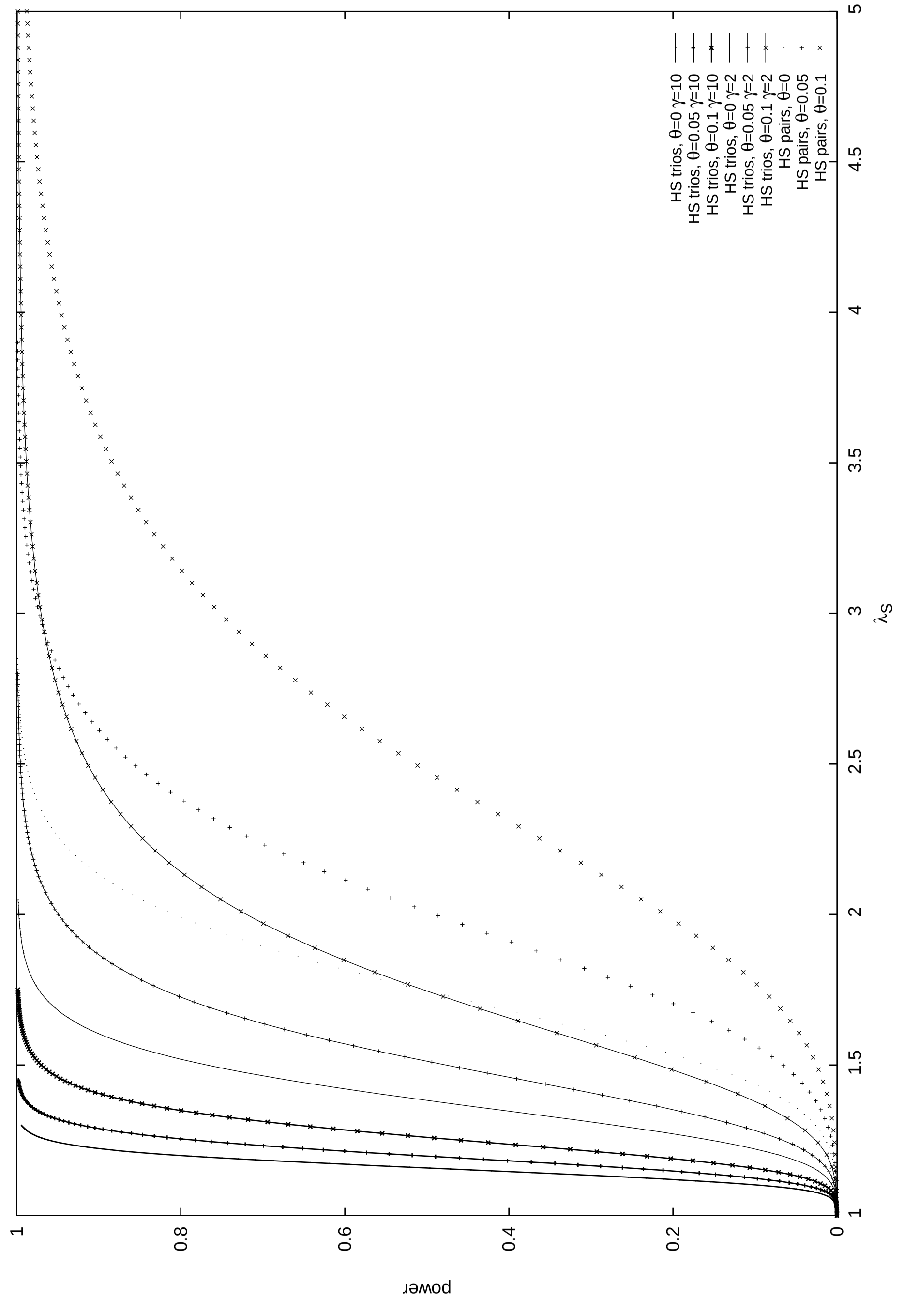
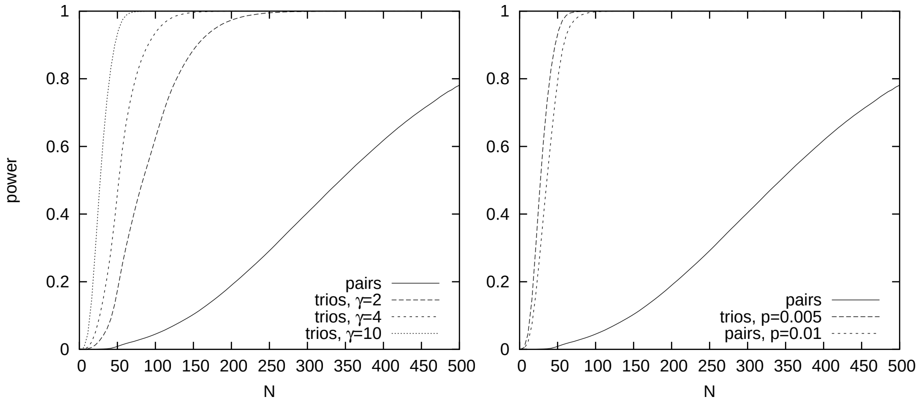
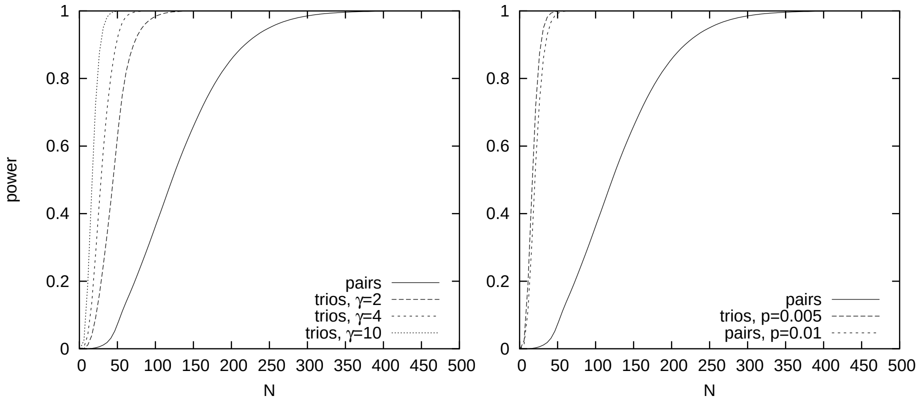


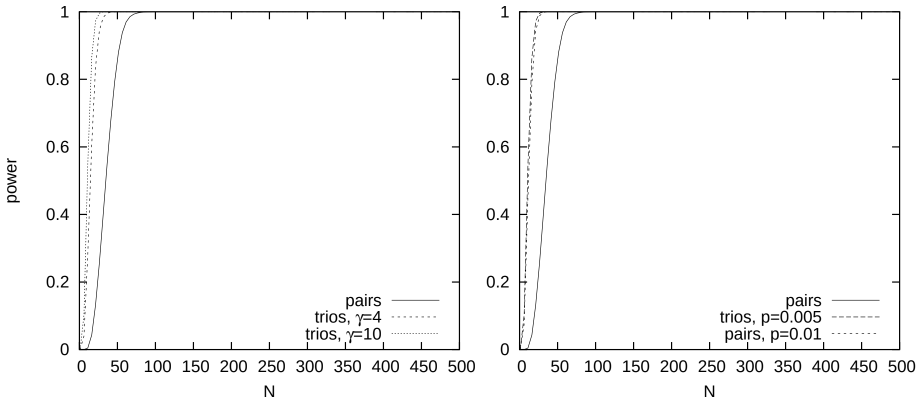
Figure 5.15: Power to detect linkage: diallelic codominant model, $\theta \geq 0$ II. 120 HS pairs or 180 HS trios



(a) NDV, $\lambda_S = 1.5$ (b) Codominant, $\lambda_S = 1.5$



(c) NDV, $\lambda_S = 2$ (d) Codominant, $\lambda_S = 2$



(e) NDV, $\lambda_S = 5$ (f) Codominant, $\lambda_S = 5$

Figure 5.16: Power to detect linkage using half sibling pairs and trios by number of pairs/trios. There are no entries for $\lambda_S = 5$ *for the NDV model,* $\gamma_1 = 2$ *because* λ_S *is limited by 3.5 at this* γ_1

Model	Power								
	50%			80%			90%		
λ_S	1.5	2	5	1.5	2	5	1.5	2	5
trios									
NDV, $\gamma = 2$	79	45	–	124	56	–	152	70	–
NDV, $\gamma = 4$	51	27	15	73	39	19	87	45	20
NDV, $\gamma = 10$	27	17	11	37	22	13	45	26	15
codom, $p = 0.01$	36	20	11	50	27	15	53	32	17
pairs									
-	336	118	33	> 500	180	44	> 500	212	51

Table 5.7: Number of HS pairs and trios required to reach 50%, 80% and 90% power. There are no entries for $\lambda_S = 5$ for the NDV model, $\gamma_1 = 2$ because λ_S is limited by 3.5 at this γ_1

ual?' (Karigl, 1982). Distributions of identity states for multiple *unaffected* relatives were examined by Thompson (1974) who proposed an algorithm based on equivalence classes for calculating the distribution among relatives. Whittemore and Halpern (1994b) extended this method, developing an algorithm which used peeling to calculated the probability of any identity state for a general pedigree. They also explored calculation of the probability of phenotype vectors for relative sets by conditioning on genotype.

In order to study the aggregation of disease among relatives and how the identity distribution varies from the null at loci linked to disease a genetic model is required to describe the relationship between genotype and disease. We have expressed this model in terms of parameters from genetic components of variance and the distribution of disease among relatives has been calculated by conditioning on identity states. We have considered only single-locus models, but extension to multilocus models along the lines outlined by Risch (1990a) would be possible.

Recurrence risk ratios (λ_R and λ_{R_1,R_2}) were used to summarise the aggregation of the disease phenotype among relatives. While λ_S depends only on three genetic parameters (the mean, additive and dominance variance of the trait), λ_{R_1,R_2} depends on many more. This also emphasises the potential of using relative trios to discriminate between genetic models that would be indistinguishable using only pairs, in particular, those parameters that depend on the third moment. Because of this, however, we restricted attention to two special case models.

As noted by Rybicki and Elston (2000) and Schliekelman and Slatkin (2002), not all values of λ_S are consistent with all genetic models. Two special case models were used in this chapter and their implications for the valid range of λ_S was explored in section 5.3. The diallelic models allowed λ_S to cover a wide range when the disease related allele was not very common, which is a reasonable assumption.

However, under the NDV model, λ_S was restricted by the skew of the genetic trait, with $\lambda_S < 1.25, 1.9, 3.5, 5.75$ for $\gamma_1 = 0, 1, 2, 3$ respectively. It is only when $\gamma_1 \geq 4$ that extended ranges for λ_S are possible. Figure 5.6 uses familiar distributions to help interpret the magnitude of γ_1. To allow even relatively moderate ranges for λ_S (eg $\lambda_S \in [1, 5]$) under the NDV model, then, we are implicitly requiring that the distribution of the genetic trait be very strongly skewed.

In particular, a zero-skew model with NDV is unlikely to be plausible for any trait for which the genetic effect is strong enough to detect since this more restricts $\lambda_S < 5/4$ and this limit is only attained for large genotype relative risks (see figure 5.5).

5.6.1 Ambiguous identity state

The results in this chapter regarding power show that under single-locus models, HS trios are likely to provide substantially more power in a linkage analysis than HS pairs. Note, however, that this work assumed IBD status could be unambiguously determined. In practice, markers are not completely polymorphic and there will be an amount of uncertainty about IBD sharing. This uncertainty can be reduced by the use of multipoint methods and by typing additional relatives. While for HS pairs three parents need to be typed, an HS trio has four parents. This increases the number of people to be typed , but HS trios will be more economical than pairs: 40 HS trios contain a total of 280 individuals while 60 HS pairs contain 300 individuals if all parents are available. Note that while typing parents will help resolve IBD status, for half siblings, who will share at most one allele IBD at any locus, typing at least the common parent will be particularly useful for resolving mispaternities. However, it has been shown that typing only affecteds in the first stage of a two-stage genome screen is a more efficient strategy (Holmans and Craddock, 1997) and this is likely to hold for HS too, if markers are sufficiently closely spaced that mispaternities may be detected.

5.6.2 Multilocus models

This chapter has focused on single-locus models. For multilocus models, additional parameters must be introduced to describe the manner in which genetic effects at different loci contribute to the overall genetic risk of disease. Risch (1990a) showed that the overall results under multilocus models were similar to single locus models, in terms of whether a particular relative pair offered more or less power than another to detect linkage, although the magnitude of that power differed. It is likely that this would apply to trios also, but further work would be needed to confirm this.

5.6.3 Relationship to other IBD scoring methods

The work in this chapter has compared the power to detect linkage using a likelihood ratio test for half sib pairs and trios. Half siblings pairs and trios can be in only one of two possible identity states and so were chosen to simplify computation of power. The statistic used would work equally well for relatives who could be in more than two identity states, assuming identity states could be identified from marker data. However, exact power calculations would be less straightforward because we would be dealing with multinomial distributions although they would still be possible to calculate numerically using simulation.

The IBD scoring tests proposed by Whittemore and Halpern (1994a) assign a score to each identity state and test whether the sum of scores over relatives differs from that expected using the statistic

$$\mathcal{T} = \frac{T - E(T)}{[V(T)]^{1/2}}$$

where T is the sum of scores and $E(T)$ and $V(T)$ its mean and variance. T can also be the calculated as the mean score conditional on observed IBD sharing among relatives, allowing the data to be used which do not completely determine the identity distribution.

In cases when IBD sharing *can* be unambiguously determined, and for relative sets with only two possible identity states, the test used in this chapter is equivalent to any IBD scoring test since both can be considered functions of the number of sets in a particular identity state. For other relatives, even when IBD sharing is unambiguous, the two tests will not be equivalent, because the test used in this chapter uses a likelihood ratio test to compare the observed sharing with that expected under a multinomial model while the IBD scoring test assigns a numerical value ('score') to each identity state allowing the data to be combined over different relatives. The multinomial test used here was chosen to provide an exact test under the theoretical case of complete identity state information to allow us to compare the relative power of different relatives in a linkage test as a function of the parameters of the genetic model. The IBD scoring test, though, is obviously more suitable to real data, although power under this test will also depend on the score chosen.

5.6.4 Framework for analysis of mixed datasets

Any dataset collected is likely to contain a mixture of different relative sets and no suggestions have been made in this chapter regarding how such data might be analysed. The Whittemore and Halpern IBD scoring method generalises easily to datasets containing different sizes and classes of relative sets.

The framework used above can also be applied to mixed datasets. Suppose n_p out of N_p pairs and n_t out of N_t trios are observed to be in id states S_5 and T_8 respectively. Lod scores are additive, so we could just add the lod scores from the pairs and trios to get an overall lod score, $\log_{10} \Lambda'$ where

$$\Lambda' = \prod_R \Lambda_R$$

and R indexes the relative types.

5.6.5 Recruitment strategies in linkage studies

Risch (1990b) showed that power to detect linkage could be expressed directly in terms of λ_R without any other genetic parameters required, assuming IBD status could be unambiguously determined close to the disease locus. Under these circumstances, more distantly related relative pairs provide greater power to detect linkage, but that increase in power falls when markers are widely spaced or are less than fully polymorphic.

The work in section 5.5 showed that, for relative trios, power to detect linkage cannot be expressed in terms of recurrence risks only and depends on the genetic model. Under a recessive model, full siblings provide considerably more power than any other relative type to detect linkage and so if the underlying model is believed to be recessive, recruitment should be targeted towards affected full siblings. Again, larger sibships are likely to be preferred.

In the case where the genetic model is not recessive, Risch (1990b) showed that affected relative pairs of higher degree can provide more power than siblings and the work in this chapter suggests that a strategy of recruiting affected HS trios would be more efficient than HS pairs or sib pairs. One could also argue that paternal half siblings from Karonga in particular offer a further advantage over full siblings when considering an infectious disease such as leprosy. As described in chapter 2, transmission is favoured by close and prolonged contact with an infected individual. In Malawi, households where the head of household has more than one wife tend to be arranged so that, within the household, children share a dwelling with their mother. This means half siblings are less likely than full siblings to share exposure. This is likely to be similar in other populations, since children will often stay with their mothers after parents split.

While recruitment of HS trios may be difficult in Western populations, among Malawian society it is common for men and women to have children by more than one partner. The number of men and women recorded by the KPS database to have had children with one or more partners is shown in table 5.8. Despite this, recruitment of affected half sibling trios may still not be feasible. Only 11 affected half sibling trios could be identified from the database and there are considerably fewer affected half sibling than full sibling pairs (99 vs 207) despite the number of (non-independent) half and full sibling pairs overall being of similar magnitude ($\sim 171,000$ and $\sim 226,000$ respectively). This difference is likely to reflect the aggregation of disease among closer relatives due to shared environmental and genetic factors.

We have not examined what increase in power might result from using other relative trios in preference to pairs, but it is likely that trios will provide more power than pairs generally. This could be examined using the theory above, but multinomial probabilities for the identity states make it computationally more complex. For this reason it is not clear whether half trios should be preferred over full sibling trios as half sibling pairs are over

Number of different co-parents	Freq	Relative Freq	Cumulative Freq
(Female co-parents per man)			
1	29,269	66.19	66.19
2	9,663	21.85	88.05
3	3,420	7.73	95.78
4	1,166	2.63	98.42
5	438	0.99	99.41
6	148	0.33	99.75
7	53	0.11	99.87
8	30	0.06	99.93
9+	27	0.06	100.00
Total	44,214	100.00	100.00
(Male co-parents per woman)			
1	47,361	83.83	83.83
2	7,569	13.39	97.23
3	1,315	2.32	99.56
4	205	0.36	99.92
5	36	0.06	99.98
6	6	0.01	100.00
Total	56,492	100.00	100.00

Table 5.8: Frequency table for the number of different co-parents for fathers and mothers

	Frequency	
Sibship size	Full siblings	Half siblings
2	165	118
3	18	11
4	2	0
5	1	0

Table 5.9: Frequency table for affected full- and half-sibships

full sibling pairs (Risch, 1990b).

Unless a method of analysis has been chosen which depends on the ASP design, it is recommended that HS pairs and above be recruited along with full sibling pairs and above for use in linkage studies. A design such as this has the additional advantage that if mispaternities are detected among full sibling recruits at genotyping, they may still be used in the analysis as half siblings. Recruitment of other relatives may be less straightforward, but, if available, multiplex pedigrees are likely to provide the greatest power in linkage studies.

5.7 Summary

The results this chapter demonstrate convincingly that larger affected relative sets can offer greater power to detect linkage than smaller ones, particularly when the genetic effect is small (as measured by λ_S). However, larger relative sets are likely to occur less frequently than smaller sets (as indicated by the availability of affected HS pairs and trios discussed in section 5.6.5). Therefore the degree to which this consideration should influence the recruitment of affected relatives for use in linkage studies needs further work before it is clear whether specific targetting of larger relative sets would be worthwhile given the extra effort that would be required to ascertain them.

CHAPTER 6.

LINKAGE ANALYSIS STUDY OF LEPROSY

6.1 Introduction

Of the three main aims introduced in section 1.1, the third was to examine whether positive results from genetic studies of leprosy in other populations could be replicated in the Karonga population. Recall from figure 2.4 that a person's genes may affect immune response and determine whether he or she develops clinical disease and/or the type of leprosy that may develop once infection has taken hold. This chapter describes the analysis of data from the KPS in a partial genome screen to test for linkage to leprosy per se and also to the type of leprosy among diagnosed cases.

The strategy for the scan is discussed in section 6.2; the rationale for the choice of selected chromosome regions under study and description of the data and methods used are presented in section 6.3 and preliminary results are reported in section 6.4. Some discussion of the methods used is given in section 6.5; a full discussion of these results, particularly in relation in published studies is made in chapter 7.

6.2 Strategy

As discussed in section 2.4.8, genotyping all available individuals across all regions of chromosome is not the most efficient strategy when conducting a genome screen. Within the KPS, blood for genotyping has already been collected from multiplex families as described in section 3.1. In this project, attention has been restricted to those families which included two or more affected members from whom blood has been collected and stored so that extraction of genotypic data was possible.

The data comprised nuclear families and extended pedigrees, and although initial ascertainment of families was based on affected sibling pairs, many pedigrees did not contain an affected sibling pair with blood collected, since some individuals could not be bled, as described in section 3.1.

It was decided to perform a two-stage genome screen as follows.

Stage I Screen selected chromosome regions among nuclear families containing an affected sibships using a coarse marker map with marker spacing about 7–10cM.

Stage II Identify regions that show potential linkage (MLS > 1) and reanalyse using a finer marker map and all nuclear families and extended pedigrees.

Most recommendations for efficient strategies (e.g. Holmans and Clayton, 1995) suggest that typing parents or unaffected siblings in the first stage of a genome scan using nuclear families is less efficient than typing additional affected sibling pairs. However, in this project, ascertaining additional affected siblings was not an option and only a relatively small number of nuclear families (91) were available. Also, it was not known how much misreporting of paternity there might be in this population. Parents (or unaffected siblings) provide additional information not only to help resolve IBD sharing but also to confirm paternity and it was decided to type parents when available. When unaffected siblings were available but one or both parents were not, it was decided to type one or two unaffected siblings respectively.

In the second stage, a particular issue which needed to be addressed is the choice of which members of the extended pedigrees should be typed. Part of an efficient strategy is to ensure individuals are not typed unnecessarily while maintaining power. It was decided that all available affected members should be typed in order to maximise power, but, while typing all unaffected members is unlikely to be optimal, it was not clear what the most efficient strategy was for their selection.

Unaffected relatives have unknown phenotype: we cannot assume someone is not susceptible to leprosy simply because they do not show clinical signs of disease. They may be infected, but have yet to develop disease, or they may not have been exposed. Therefore, whichever method of linkage analysis is used, unaffected members will not contribute directly to the statistic used to detect linkage, but can provide information that helps to infer the IBD sharing configuration more accurately within the set of affecteds.

Simulation was used to compare the expected information about IBD sharing under different selection strategies, and to select unaffected pedigree members for genotyping on the basis of these results. The development of the method used is discussed further in section 6.3.4.

6.3 Data and methods

6.3.1 Genotyping and choice of regions for genome screen

Genotyping was carried out by Drs Jodene Fitness and Branwen Hennig in Professor Adrian Hill's laboratory at the Wellcome Trust Genome Centre in Oxford.

145 markers were typed in regions on chromosomes 5, 6, 9, 10, 15, 20, 21 and X (see appendix D), at an average spacing of 7cM. Choice of these regions was based on evidence described in section 2.7.5. Chromosome 6 contains the MHC region and the 6q25–q27 region recently identified in a Vietnamese population (Mira et al., 2003). Chromosome 9 contains a cytokine gene cluster and the markers are in the same panel as markers in the region of chromosome 10p13, which was shown, together with a region on chromosome 20p12, to be linked to leprosy in a linkage analysis in South India. Recent work on susceptibility to TB implicated the X chromosome in a study in the Gambia (Bellamy et al., 2000) and unpublished work from the same study found evidence of linkage to chromosome 15. These are both being included in case they are linked to general mycobacterial susceptibility. The chromosome 21 region is short, but contains an important cytokine gene cluster including interferon and IL10 receptors which are very important in host defence. Genotyping errors and genetic relationships were checked using pedcheck (O'Connell and Weeks, 1998).

6.3.2 Nuclear family data

The identification of and collection of blood from affected sibships was described in section 3.1. A total of 91 affected sibships with blood collected were identified. Blood had not been collected, however, from all their parents. Where one or both parents were unavailable, one or two unaffected siblings respectively (if available) were included to help resolve IBD status.

Family size	# affected sibs	# parents	# unaffected sibs	Frequency
2	2	0	0	10
3	2	1	0	4
3	2	0	1	9
4	2	2	0	15
4	2	1	1	23
4	2	0	2	15
4	3	1	0	1
4	3	0	1	2
5	3	1	1	1
6	4	2	0	2
Number of nuclear families				82
Number of affected sib pairs				100
Number of independent affected sib pairs				90

Table 6.1: Breakdown of affected sibships by size and number of parents and unaffected siblings typed

Examination of the genetic data with the help of pedcheck (O'Connell and Weeks, 1998) indicated that ten families were likely to contain mispaternities and in one further family genotypic data indicated that one of the affected siblings was not related to his purported mother or brother. In two of the mispaternity families, the children appeared to share their father (though not the father they claimed to share), and were included in the analysis as if the father were not typed. The other nine (the seven mispaternity families and the one family with an apparently unrelated affected child) were excluded from this analysis. Although this is a relatively small sample, this level of misreporting of parents (9/91=0.099%) is comparable to that in Western populations.

Table 6.1 shows the breakdown of these sibships after exclusions by size and availability of parents and unaffected siblings. In total, there were 336 individuals to be typed.

6.3.3 Extended pedigree data

Extended pedigrees were identified for analysis according to the following protocol:

1. identify all affected individuals from whom blood has been collected - they form the set $\mathcal{A}$
2. for each $a \in \mathcal{A}$, identify all $a' \in \mathcal{A}$ who are first, second and third degree relatives of a. Together they form the set $\mathcal{R}_a$. If $\mathcal{R}_a$ exactly equals $\mathcal{R}_b$ for a previously identified b, discard $\mathcal{R}_a$
3. identify any sets $\mathcal{R}_a, \mathcal{R}_b$ which have a non-empty intersection and for which r_a is connected to r_b for all $r_a \in \mathcal{R}_a; r_b \in \mathcal{R}_b$ and replace with a single set $\mathcal{R}_a \cup \mathcal{R}_b$
4. for each $\mathcal{R}_a$, construct a pedigree by adding the following:
    (a) unaffected first, second and third degree relatives of $\mathcal{R}_a$ from whom blood has been collected
    (b) any unbled relatives linking already selected pedigree members (needed to create a connected pedigree)

185 pedigrees were identified in this way. These included 11 nuclear families from the sib pair study who had no other affected and bled relatives, 71 larger pedigrees that contained sib pairs from the stage I dataset and 103 extended pedigrees that did not contain affected sibships.

6.3.4 Selection of unaffected pedigree members

As discussed above, simulation methods were used to compare different selection strategies for unaffected pedigree members. For any given pedigree, we can divide the bled members of a pedigree into a set of affecteds, $\mathcal{R}$, and a set of unaffecteds, $\mathcal{R}'$. One way to compare different combinations of unaffected relatives is to compare the expected information about IBD sharing among $\mathcal{R}$ at a locus when simulated data are analysed using the same set $\mathcal{R}$ and with different subsets of $\mathcal{R}'$. This will provide a measure of expected information under the null hypothesis.

In this section, the construction of such a method and its application to extended pedigree data from Karonga district is described. Guidelines for the choice of unaffected members in future studies are discussed in section 6.5.1.

Measures of Information

Various methods to calculate the expected information for linkage are described by Nicolae (1999); the two most commonly used, entropy and variance of the IBD configuration, are described here. In order to calculate the expected information, we ideally want to use a per-family measure of information, I_i, such that the information in a sample of families can be expressed as a weighted sum, $I = \frac{\sum_i w_i I_i}{\sum_i w_i}$ over each family, i. This allows straightforward calculation of a measure of expected information. The weights may depend on the shape of the pedigree, but in a sample of (simulated) pedigrees with the same shape and characteristics, $w_i = w_j \; \forall \; i, j$ and $\bar{I} = 1/n \sum I_i$ is the expected per family information.

Entropy The residual uncertainty in a probability distribution can be measured by its entropy, $E = -\sum_i P_i \log_2 P_i$ where P_i is the probability of the ith outcome (Shannon, 1984). Kruglyak and Lander (1995) used this definition to describe the information content of an inheritance distribution at a locus x by

$$I_E(x) = 1 - \frac{E(x)}{E_0}$$

where $E(x)$ is the entropy of the inheritance distribution and E_0 is the entropy in the absence of genotype data. Note that in this case, the probability distribution is uniform over all 2^{2n-f} equivalence classes of inheritance vectors, and so $E_0 = 2n - f$ bits. When the inheritance vector is known with probability 1 (eg at a fully informative marker), $E = \log_2 1 = 0$, (and $I_E = 1$) so

$$0 \leq I_E(x) \leq 1$$

and $I_E(x)$ approaches 1 as information increases.

Variance of the IBD configuration When IBD scoring methods are used to detect linkage, the score is averaged over all IBD configurations consistent with observed data, weighted by the probability of these configurations (given the data). If Z is the IBD configuration at a locus, the variance of Z is a measure of the certainty with which Z is known - the lower var(Z), the less uncertainty, and so the more information.

This leads to a natural definition of information as

$$I_V = 1 - \text{var}(Z)$$

This is equal to 1 if and only if $\text{var}(Z) = 0$ (ie Z is known with certainty), but is not bounded below so

$$I_V \leq 1$$

and, again, I_V approaches 1 as information increases.

I_E does not depend on the method that will be used to test for linkage. Often though, it would be preferable to use a measure that is tied to the testing procedure to be used, and I_V describes the information available under the particular testing procedure. I_V has been used in this study because it is closely tied to the IBD scoring method that will be used in the linkage analysis.

Method of Simulation

We can use simulation to decide which unaffected pedigree members to include as follows:

1. choose a 'template' pedigree
2. simulate multi-point marker data (unlinked to disease) for a number copies of this pedigree and designate individuals as affected, unaffected or unavailable according to the template
3. calculate the mean per-family information, I_V, at particular loci using all affected and different combinations of unaffected pedigree members
4. compare the expected information under each combination

I have chosen to use simulate data along 3 stretches of chromosome, each with the same marker density (2cM, the minimum marker spacing we expect to use in the second stage screen) but with differing marker heterozygosity: 0.67, 0.8 and 0.9 (using 3, 5 or 10 equifrequent markers). These values for the heterozygosity were chosen because they are close to the lowest, average and highest heterozygosities for the markers which used in the first stage of the genome screen (see section 6.3.1 and appendix D).

It was decided to simulate 1,000 copies of each pedigree initially, and then simulate further copies (in blocks of 1,000) until the standard error of the mean information, se($\bar{I}_V$) was within 1% of $\bar{I}_V$. For larger pedigrees ($\geq$ 15 bits), where the 1% target was impractical due to computing time, a 2% target was set. A schematic diagram of this method is shown in figure 6.1.

Improving the speed of simulation

The simulation part of this method is very quick. However, for all but the most simple pedigrees, the number of possible combinations of unaffecteds can be very large and the analysis of the simulated pedigrees can be very slow, particularly for larger pedigrees. This can be reduced by choosing some members to be 'always typed' and by assigning the other members to *equivalence classes* and including only those combinations that contain different combinations of classes. It is easiest to define an equivalence class as used here by means of a simple example.

Consider the pedigree shown in figure 6.2 and note that the mother is unavailable for genotyping. The father is definitely useful for determining IBD, so designate him (and the affected siblings) as 'always typed'. Members 5 and 6 are both unaffected siblings. With no information about their genotype, they may be expected to contribute equally to any linkage analysis and so are considered *equivalent* in this context. Thus the combinations that would be typed in this pedigree are {1,3,4}, {1,3,4,5}, {1,3,4,5,6}. The combination {1,3,4,6} would be discarded because members 5 and 6 are equivalent, ie $E(I_V(\{1,3,4,5\})) = E(I_V(\{1,3,4,6\}))$, and combinations that did not include the father (such as {3,4,5}) would be discarded because he should always be typed. This reduces the number of combinations from eight to three.

For more complicated pedigrees, a further reduction is possible if we deem some individuals to be 'not necessary' when particular others are typed. Suppose, in the same example, that the mother were available. It might be realistic to assume that once both parents were typed, no additional information could be found by typing the unaffected siblings.

Even so, the number of combinations can remain large for pedigrees with several unaffected members and this slows the program because of the need to analyse each and every combination; the time this takes rises

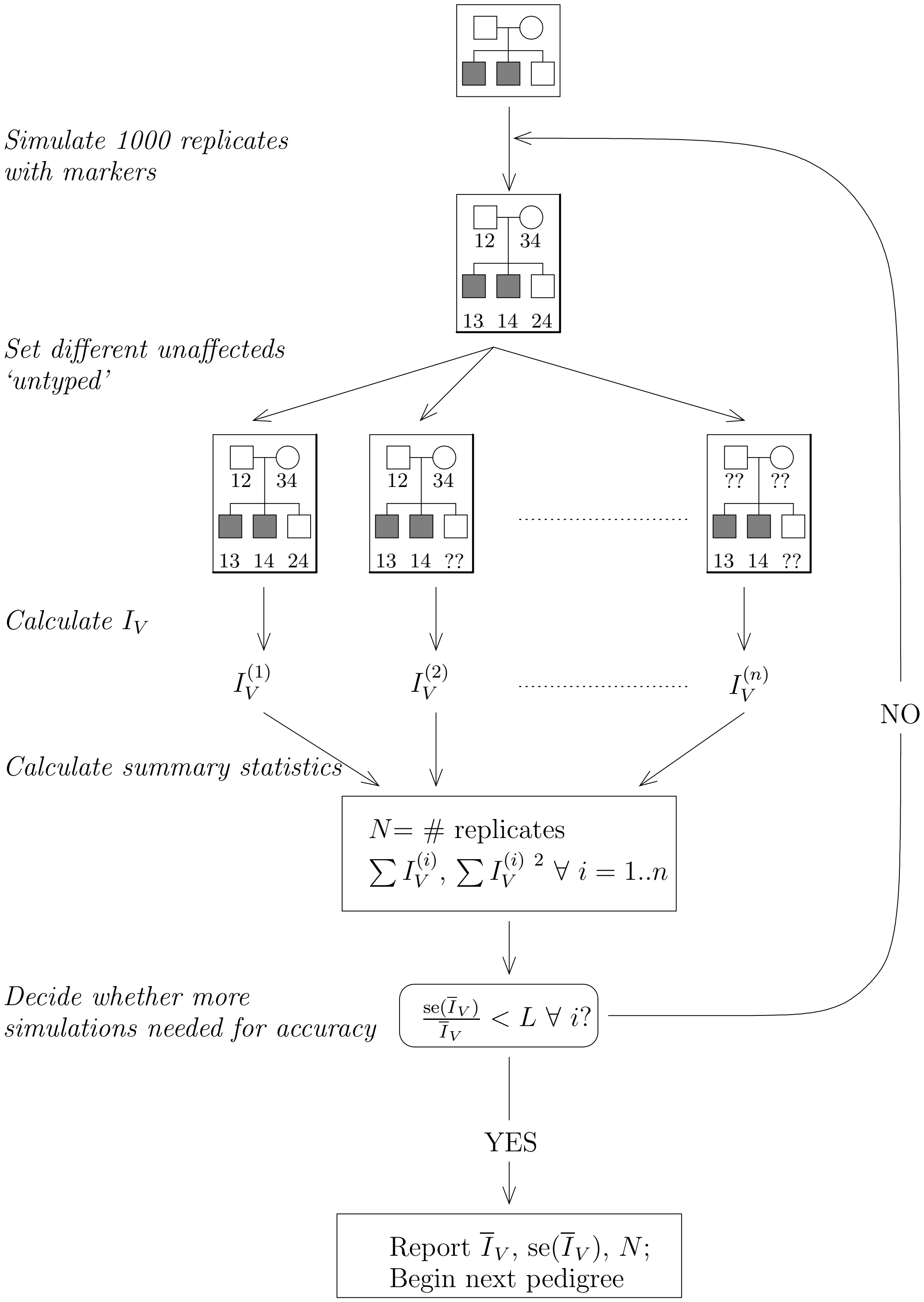

Simulate 1000 replicates with markers
12
34
13
14
24
Set different unaffecteds 'untyped'
12
34
13
14
24
12
34
13
14
??
??
??
13
14
??
Calculate I_V
$I_V^{(1)}$
$I_V^{(2)}$
$I_V^{(n)}$
Calculate summary statistics
$N=$ # replicates
$\sum I_V^{(i)}, \sum {I_V^{(i)}}^2 \; \forall \, i = 1..n$
Decide whether more simulations needed for accuracy
$\frac{\mathrm{se}(\overline{I}_V)}{\overline{I}_V} < L \; \forall \, i?$
NO
YES
Report $\overline{I}_V$, $\mathrm{se}(\overline{I}_V)$, N; Begin next pedigree

Figure 6.1: Schematic diagram of method of simulation. L is determined by pedigree size: $L = 1\%$ if pedigree smaller than 15 bits; $L = 2\%$ otherwise

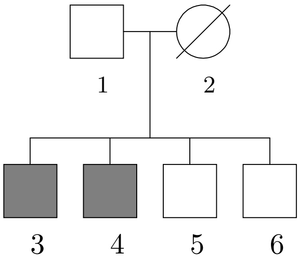


Figure 6.2: Example pedigree: nuclear family containing two affected and two unaffected siblings, with the mother unavailable for genotyping

exponentially with the number of pedigree members and linearly with the number of replicates requested.

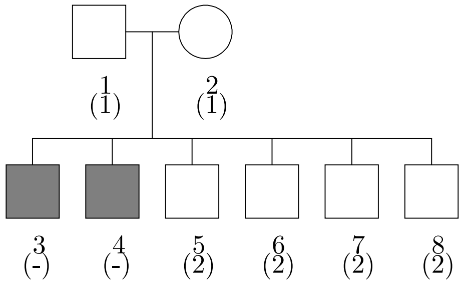
A program was written to implement this method, which takes each pedigree in turn and generates the combinations to be typed. It then calls SIMULATE (Terwilliger et al., 1993) to simulate genotype data and Allegro (Gudbjartsson et al., 2000) to analyse the data and calculates $\overline{I}_V$. It repeats this until se($\overline{I}_V$) is within the required limit for all sets i, reports results and begins again with the next pedigree.

Template pedigrees

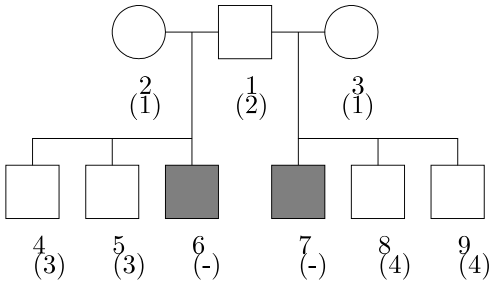
In order to examine the method for selection of unaffected pedigree members, and check that results from the program were sensible (as a means of program testing), two simple 'template pedigrees' were chosen (shown in figure 6.3), containing an affected full- and half-sibling pair and other unaffected members. The results from these pedigrees will provide a useful reference for analysing the extended pedigrees from the KPS since many of these pedigrees consist of an affected full or half sibling pair and unaffected relatives.

In addition to the three simulated stretches of chromosome that will be used for the extended pedigrees, we will use a further three stretches in the template pedigrees with the same heterozygosities, but a coarser marker spacing of 7cM (the average marker spacing used in the first stage of the screen).

Only the affected members were designated 'always typed', and the unaffected members were assigned to equivalence classes - these are also shown in figure 6.3.



(a) Affected sibling pair



(b) Affected half sibling pair

Figure 6.3: Template pedigrees used as testing examples for the simulation method. The first data line contains each individual's id number, the second the equivalence class to which they have been assigned (in parentheses)

6.3.5 Analysis of linkage to leprosy per se

Methods of linkage analysis were discussed in section 2.4; model-free methods will be used here. Although model-based methods are always at least as powerful as model-free when the true model is known, the latter are preferred in this project because

- the true model for leprosy susceptibility is not known. We could choose to test several models as suggested by MacLean et al. (1993), but because age-specific penetrances are hard to estimate due to the changing age distribution of incidence in Karonga and because many other covariates have a significant effect, there would have to be very many models considered. This would in turn lessen power because multiple testing would have to be accounted for.

- parametric methods consider individuals to be affected, unaffected or unknown and all whose disease states are known contribute to the likelihood. In nonparametric methods, unaffected individuals may be used to infer IBD sharing, but only affected individuals contribute to the likelihood. This is particularly useful for infectious disease, since the trait we are trying to model is *susceptibility* to disease. We know someone is susceptible if they have disease, but the opposite is not true: we cannot infer that unaffected people are not susceptible. Therefore it makes sense to chose a method according to which only those known to have the trait are included in the likelihood.

Stage I

It was decided to use Risch's MLS statistic in this analysis. With 82 sibships, we expect to have only fairly low power to detect linkage to a locus of moderate effect. In chapter 4, λ_S in this population was estimated to be around 2. Risch (1990b) showed that power to detect linkage (with a stringent lod score threshold of 3.0) or regions worth further investigation (with a less stringent lod score threshold of 1.0) is about 30% and 90% respectively for 100 sibling pairs at this value of λ_S. Further, it is likely that the true λ_S is lower than the estimate of 2 and is not due to the effect of a single locus, but of multiple loci. Therefore power will be even lower.

With 82 sibships and a low MLS cutoff, we would expect to be able to detect regions worthy of further investigation, but we will not have power to detect linkage according to the strict criteria that are required for a genome wide screen. Nevertheless, this exercise is a useful first stage analysis, because any areas worth further investigation can be studied in the extended pedigrees from the same population.

The sibships were analysed using MAPMAKER/SIBS (Kruglyak and Lander, 1995), within the possible triangle restrictions (Holmans, 1993). Multiple sibships (of size n) were treated by splitting them into all possible pairs and weighting by $2/n$ to account for non-independent observations.

Analysis was also conducted using the maximum likelihood binomial (MLB) method (Abel and Müller-Myhsok, 1998; Abel et al., 1998a) - although this method is only suitable for autosomal chromosomes and nuclear families, the authors claim high power and very accurate type I error rates. A modified version of genehunter which calculated this test statistic was used (Laurent Abel, personal communication).

Stage II

Methods for the linkage analysis of extended multiplex pedigrees were discussed in section 2.4.4. The allele sharing scoring statistics of Whittemore and Halpern (1994a) were used here, as implemented in Allegro, version 1.1 (Gudbjartsson et al., 2000). The exponential model of Kong and Cox (1997a) will be used to create lod scores to test the significance of the results. Both $\mathrm{NPL}_{\mathrm{pairs}}$ and $\mathrm{NPL}_{\mathrm{all}}$ will be used, since they are sensitive to different kinds of variation in IBD sharing among affected relatives - $\mathrm{NPL}_{\mathrm{pairs}}$ is more sensitive to dominant inheritance and $\mathrm{NPL}_{\mathrm{all}}$ to recessive inheritance.

Power will be higher than in stage I, but it is unlikely to be high enough to detect linkage to genes of moderate effect using the strict criteria required for genome screens. Rather, results from this stage may be used to further target regions for future study.

6.3.6 Exclusion mapping

Model-free exclusion mapping was described in section 2.4.6. Briefly, fixing λ_S under the assumption of no dominance variance fixes the allele sharing probabilities, and the likelihood of the observed data given these probabilities can be compared to the likelihood under the hypothesis of no linkage. If the resulting lod score is under -2, the region is generally ‘excluded’ - ie there is sufficient evidence against a disease associated gene responsible for a locus-specific effect of the fixed λ_S or more.

6.3.7 Analysis of linkage to leprosy type

One method to analyse linkage to leprosy type is subtype analysis, in which phenotypically defined subgroups of families are analysed separately and the results compared. This type of analysis can be applied to affected sibpair data, but is less readily applied to extended family data. Subgroup analysis was performed using both the MLS and MLB statistics, with sibpairs categorised as MB only, PB only or MB/PB. Subdividing a dataset, however, leads to lower numbers and hence lower power.

An alternative method is available. Holmans (2002) describes a method for detecting gene-gene or gene-environment interactions, which is a development of a methods proposed by Rice et al. (1999). The probability an affected relative pair share an allele IBD is regressed on the covariate of interest, which may be IBD sharing at another locus or some other covariate known to affect disease. Let p be the probability an affected relative pair share the allele they inherit from a given parent or ancestor IBD. Then the probability a sib pair share 0, 1 or 2 alleles IBD are $z_0 = (1-p)^2$, $z_1 = 2p(1-p)$, $z_2 = p^2$ and the probability any other relative pair share 0, 1 or 2 alleles IBD is $z_0 = (1-p)$, $z_1 = p$, $Z_2 = 0$. We can test for linkage using

$$LR = \sum_i \sum_{j=0}^{2} \frac{z_j \hat{f}_{ij}}{f_j}$$

where f_j and $\hat{f}_{ij}$ are the prior and posterior probabilities (given marker data respectively) that sib pair i shares j alleles IBD at the locus being tested. This represents an independence assumption for the IBD status of the maternal and paternal alleles and so implies that two disease alleles at the locus act multiplicatively. This assumption can be relaxed, however, by using robust estimates of standard errors. p can be expressed as a logistic function conditional on x,

$$p = \frac{e^{\alpha+\beta(x-\overline{x})}}{1+e^{\alpha+\beta(x-\overline{x})}},$$

where x may be the proportion of alleles shared at a locus known to affect disease (in the case where we want to test for linkage at one locus while controlling for the effect of another locus or test for interaction between the two loci) or a covariate known to affect disease. Then

$$T = 2\log\left(\frac{LR(\hat{\alpha},\hat{\beta})}{LR(\alpha=0,\beta=0)}\right)$$

is a test of linkage to the locus allowing for interaction with x, and

$$S = 2\log\left(\frac{LR(\hat{\alpha},\hat{\beta})}{LR(\hat{\alpha},\beta=0)}\right)$$

is a test of interaction between the new locus and x.

The latter test is used in this project to test for linkage to leprosy type within the nuclear family data. Here x is the pairwise type of leprosy in an affected sib pair (which is a categorical variable with three levels: MB/MB, PB/MB and PB/PB) and we use robust estimates of standard errors, allowing for clustering within sibships (as described in section 4.5.3). Hence we are testing whether the probability that two relatives inherit the same allele IBD varies according to the pairwise leprosy type. Out of the 83 sibships in the dataset, there are only two MB/MB pairs - too few to analyse as a single group. Instead, pairs will be analysed as type-concordant and type-discordant, with and without the MB/MB pairs. The stata package ibdreg (Clayton, 2002) will be used to perform the analysis.

6.4 Results

6.4.1 Selection of unaffected members from extended pedigrees

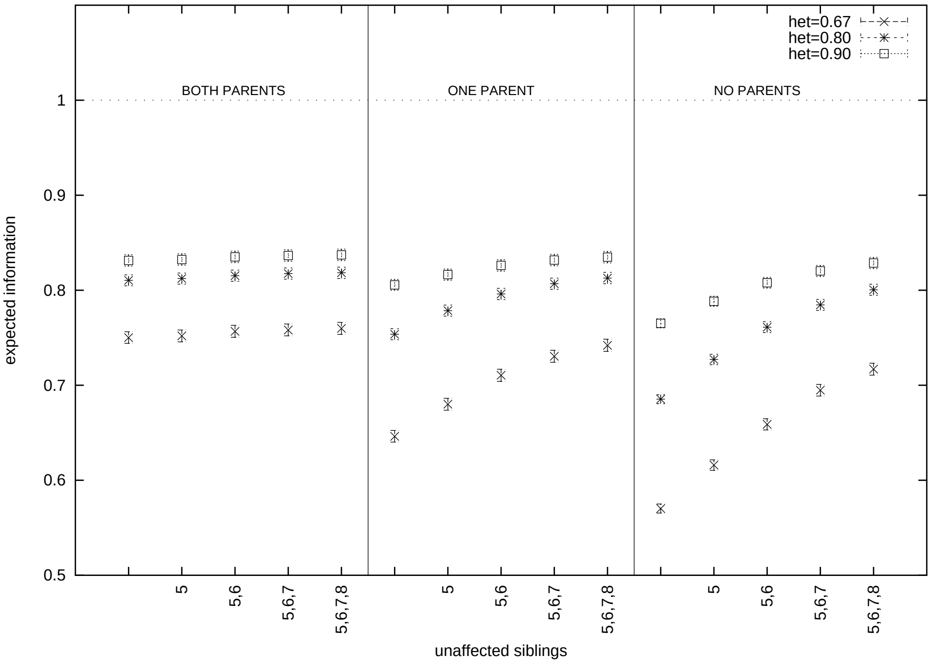
Template pedigrees

The number of simulations required to reach the target se($\overline{I}_V$) < $\overline{I}_V/100$ was small (4,000 for each template). The size of each pedigree and the number of simulations required to reach the 1% target is shown in table 6.2. The expected information for each pedigree and under each combination is shown in figures 6.4 and 6.5.

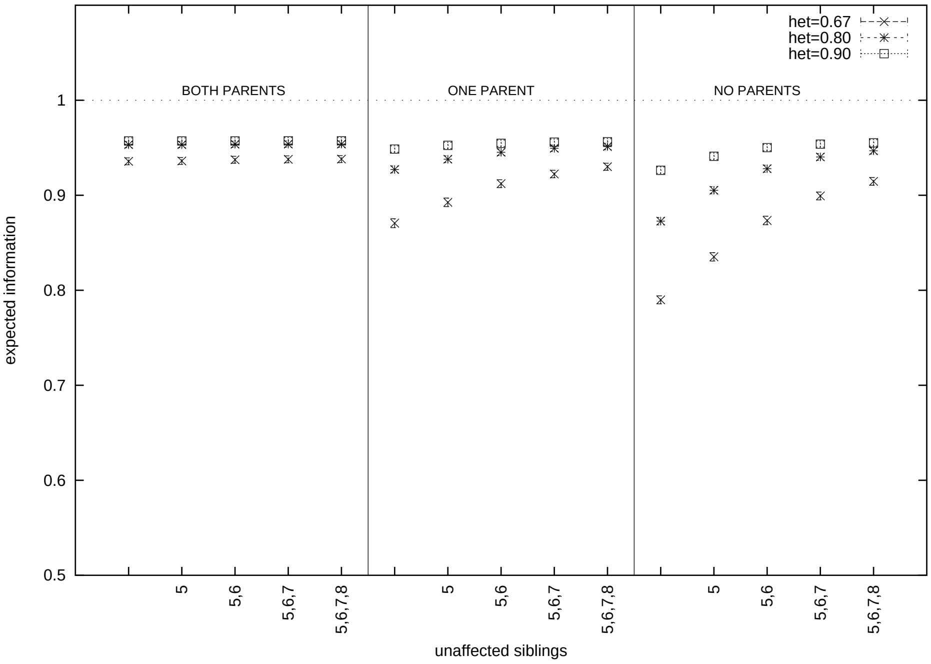
Pedigree	Size ($2n - f$)	# Simulations	# Combinations
(a) Sibling pair	10	4,000	15
(b) Half sibling pair	9	4,000	42

Table 6.2: Size of template pedigrees and maximum number of simulations per combination required to reach the target se($\overline{I}_V$) < $\overline{I}_V/100$

The results of the analysis of the template pedigrees are logical and reasonable, supporting the validity of the method. They lead to simple

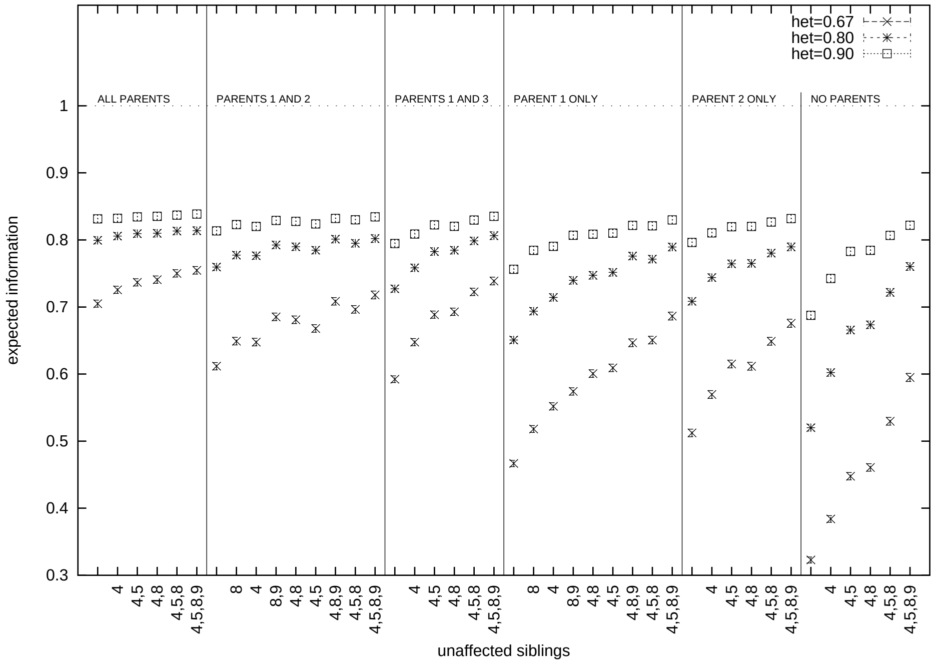


(a) Coarse marker map (marker spacing=7cM)

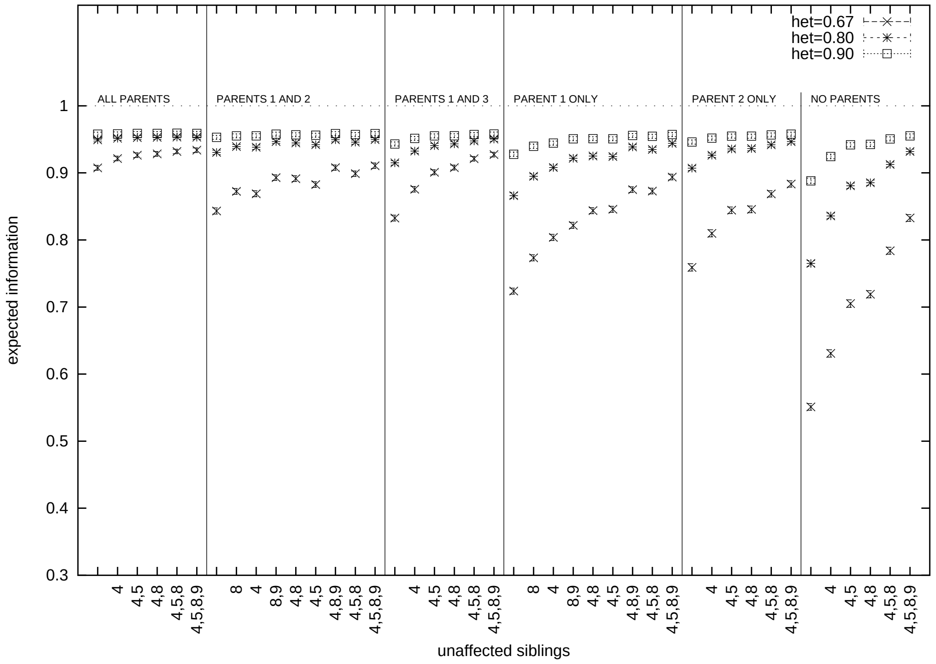


(b) Fine marker map (marker spacing=2cM)

Figure 6.4: Expected information about IBD sharing (I): results from simulations using the affected sibling pair template pedigree. The expected information and 95% confidence interval are shown for each combination of unaffected members



(a) Coarse marker map (marker spacing=7cM)



(b) Fine marker map (marker spacing=2cM)

Figure 6.5: Expected information about IBD sharing (II): results from simulations using the affected half-sibling pair template pedigree. The expected information and 95% confidence interval are shown for each combination of unaffected members

Members available	Selection of unaffected relatives
Affected sibling pair	
Both parents	both parents, no unaffected siblings
One parent	one parent, two unaffected siblings
No parents	four unaffected siblings
Affected half sibling pair	
All parents	parents, no unaffected siblings
CP[a]; 1 AP[b]	parents; 2 unaffected siblings (from missing parent side)
CP only	parent; 3 unaffected siblings
Both APs only	parents; 3 unaffected siblings
1 AP only	parent; 4 unaffected siblings
no parents	4 unaffected siblings

Table 6.3: Rules for selection of unaffected pedigree members in future linkage studies: affected sibling and half-sibling pairs. [a]CP denotes common parent; [b]AP denotes alternate (not common) parent

rules for selection of unaffected pedigree members in future studies, shown in table 6.3 for a coarse (7cM) map with 80% marker heterozygosity.

As might be expected, typing parents provides most information. If all parents are available, unaffected siblings provide no significant increase in information, except for half-siblings in the case when a coarse marker map with low heterozygosities is used, and even then the increase is fairly small.

When one or both parents are not available, the expected information can be considerably lower, and the number of unaffected siblings required to replace this lost information is often high, particularly at wider marker spacing and lower heterozygosity. Consider a genome screen of affected sibling pairs using a 7cM spaced marker map. If marker heterozygosity is about the average it was in stage I of our screen (~ 0.8), then if no parents are available the expected information would be significantly lower, even typing four unaffected siblings, than if all were parents available. The effect is similar though less pronounced when a fine map is used. This emphasises the potential increase in power that can be provided if parents are typed.

Extended pedigrees from Karonga data

Ninety extended pedigrees did not need to be simulated to determine the 'best subset'. This was either because it was clear that all unaffecteds in a pedigree should be typed, or because the pedigree matched one of the template pedigrees already studied.

This left 79 pedigrees to be examined using simulation. Their members fall into three categories:

- affected members;
- unaffected members who clearly belonged in any 'best subset' (eg parents of an affected sibling pair);
- unaffected members whose inclusion in any 'best subset' was unclear, hereafter termed *spares*.

There were a total of 219 spares, and the number per pedigree ranged from one to seven.

The 'best subset' of unaffected pedigrees was chosen according to the following protocol. For any pedigree, let C_i denote the ith combination when all n combinations are ranked according to the expected per-family information (so $I_V(C_1) > I_V(C_2) > \cdots > I_V(C_n)$) and let $\#C_i$ denote the number of individuals to be typed in combination C_i. Then eliminate those combinations C_j where $I_V(C_j)$ is significantly less than $I_V(C_1)$ (at the 5% level). Now pick the subset with the least number of individuals to be typed. If there is a tie, choose the combination with the highest I_V from the set in the tie. This protocol means the chosen subset will require the least genotyping while ensuring power is not significantly below the maximum possible.

After applying this procedure, 110 spares were included in the 'best subsets', and 109 were discarded. The average saving per pedigree, by the number of spares available, is shown in figure 6.6. A common measure of pedigree size is $2n - f$ bits, where n represents the number of non-founders and f the number of founders in the pedigree. The sizes of pedigrees before and after spares were discarded are shown in histograms in figure 6.7.

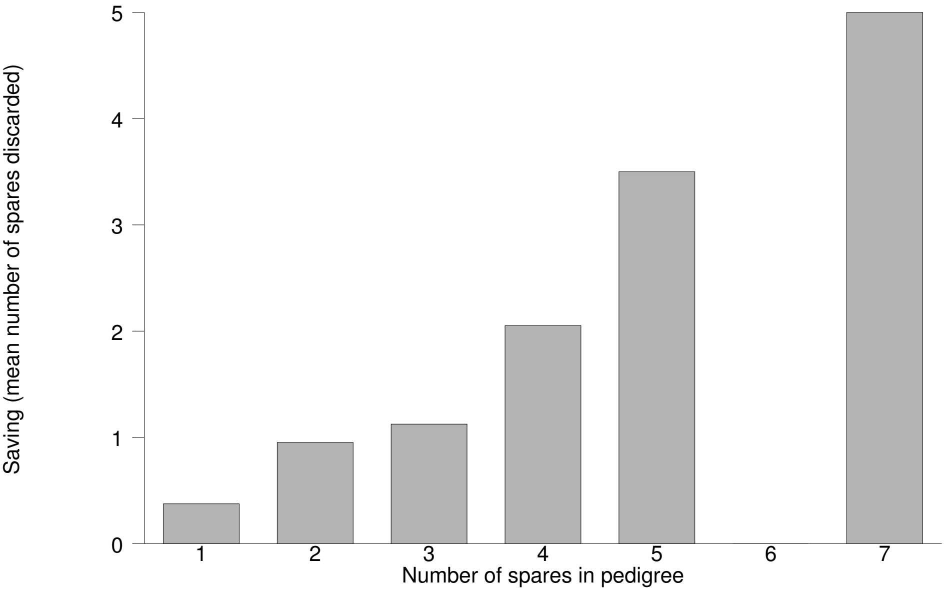


Figure 6.6: Mean number of unaffected pedigree members deemed to be not necessary to type after simulations. Note: no pedigree contained exactly six spares

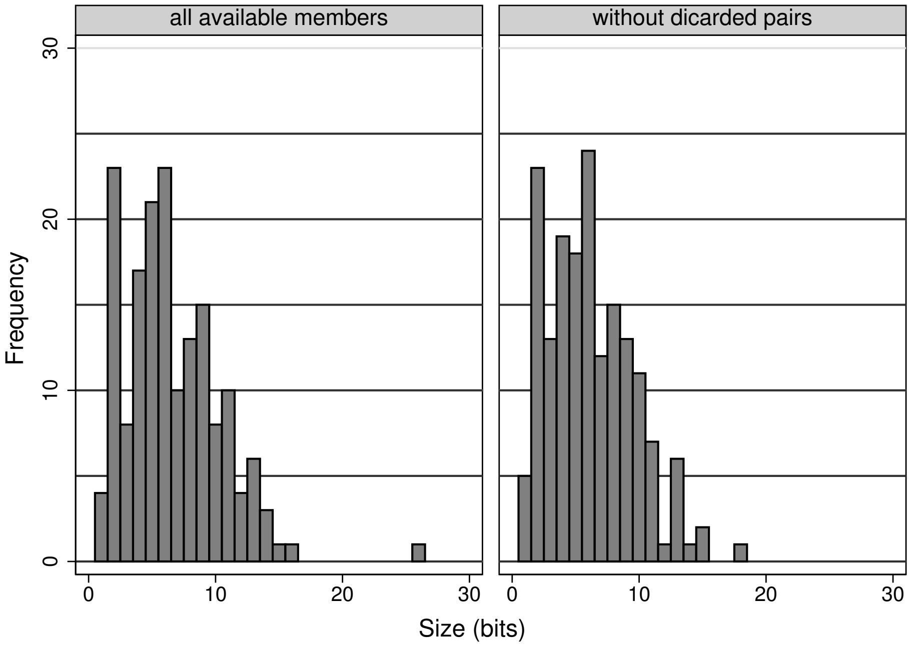


Figure 6.7: Histogram showing sizes of the extended pedigrees in bits

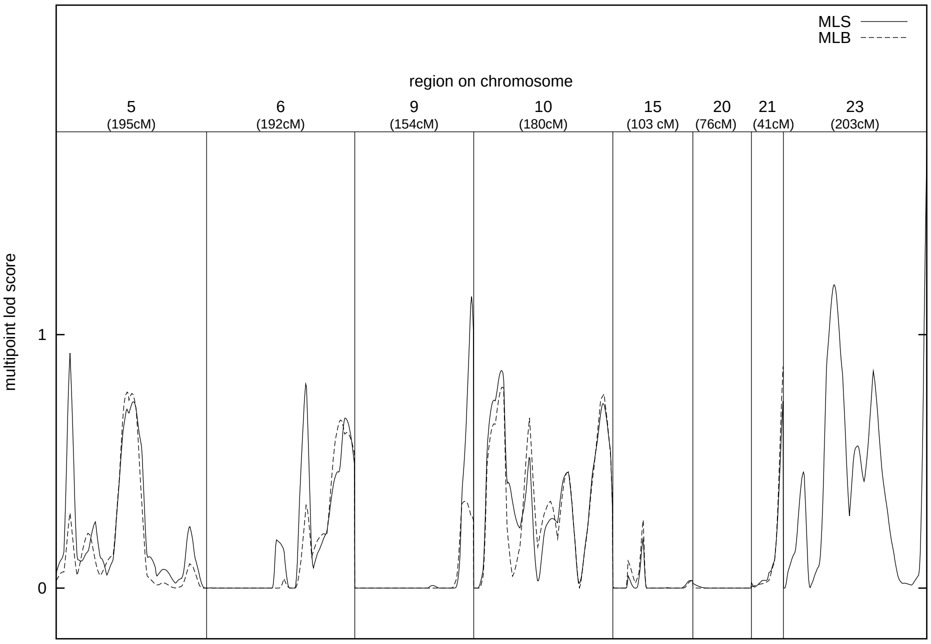


Figure 6.8: Initial results from stage I scan: lod score curves for linkage analysis of leprosy per se among affected sibships. Note: the MLB method is not suitable for the X chromosome

6.4.2 Linkage to leprosy per se

Stage I

Mega2 (Mukhopadhyay et al., 1999, 2001) was used to prepare data for the different analyses. The position of the markers used in this first stage analysis, statistics for each marker including the proportion of individuals successfully typed and marker polymorphism are given in appendix D, together with single-point MLS scores for linkage to leprosy per se. Four markers exceeded the criteria for suggestive linkage: D21S266 (MLS=1.39), DXS993 (MLS=1.06), DXS1073 (MLS=2.09) and D5S644 (MLS=1.17). Multipoint MLS and MLB curves are shown in figure 6.8.

The results of the multipoint MLS and MLB analysis are broadly similar. There are two ‘spikes’ that appear on the MLS results but not the MLB results: at the beginning of the chromosome 5 region and at the end of the chromosome 9 regions. In particular, the MLS spike on chromosome 9 reaches 1.1. The MLS also exceeded 1 in a region chromosome Xp11 between marker DXS993 and DXS991. Additionally, both the MLB and

MLS scores approached 1 in two regions on chromosome 10 and at the end of the chromosome 21 region. Of particular interest was the first of these peaks on chromosome 10 which coincided with that found in the Indian scan (Siddiqui et al., 2001).

The high singlepoint score for D5S644 was not supported by the multipoint analysis and the multipoint MLS peak on chromosome 9 was not supported by the singlepoint analysis. It was also decided to discount the very high score at the end of chromosome X (DXS1073), since it was not supported by any neighbouring markers.

Stage II

It was decided to focus on the regions located on chromosomes 10, 21 and X which showed suggestive evidence for linkage in the sib pair study and that markers in these regions should be typed in the extended families. To date, additional chromosome 21 markers have been typed in both the nuclear families and extended pedigrees, while additional chromosome 10 markers have been typed in only the nuclear families and no additional chromosome X markers have yet been typed.

Chromosome 10 followup

After typing 14 additional chromosome 10 markers around the 10p13 region, the MLS and MLB scores fell in the 10p13 region (see figure 6.9). In the Indian study which first implicated the 10p13 region (Siddiqui et al., 2001), the majority of the cases were PB leprosy and supporting evidence for this locus was found in the Vietnamese study only among PB cases. It was decided to analyse PB concordant, MB concordant and discordant sibpairs from the KPS data separately to examine whether the same pattern was found in Karonga. The results of these analyses for the initial markers only and all markers are shown in figure 6.10. There was evidence for linkage around the 10p13 region among PB concordant pairs only, but after the stage II markers were typed, this peak localised to the 10p14 region, adjacent to that identified in the Indian scan.

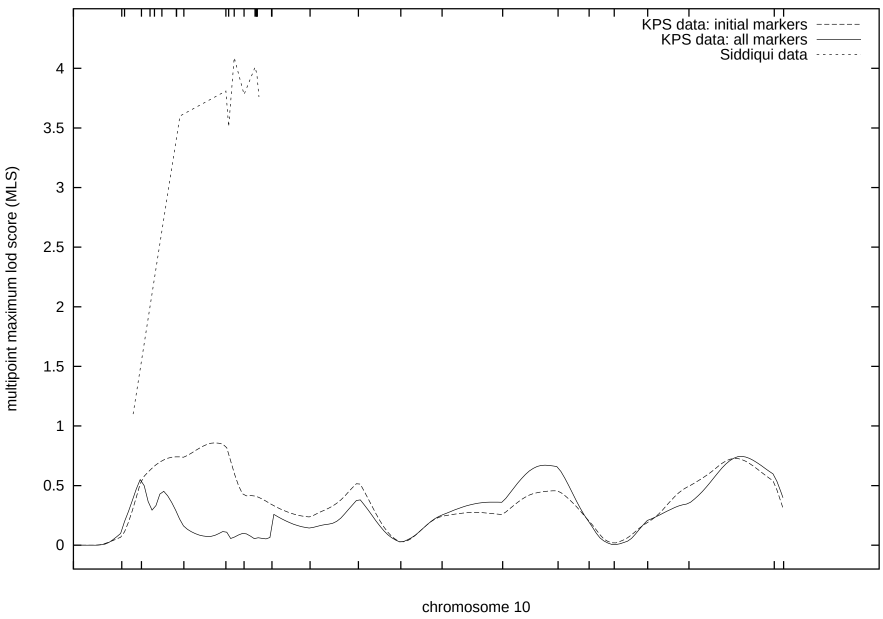
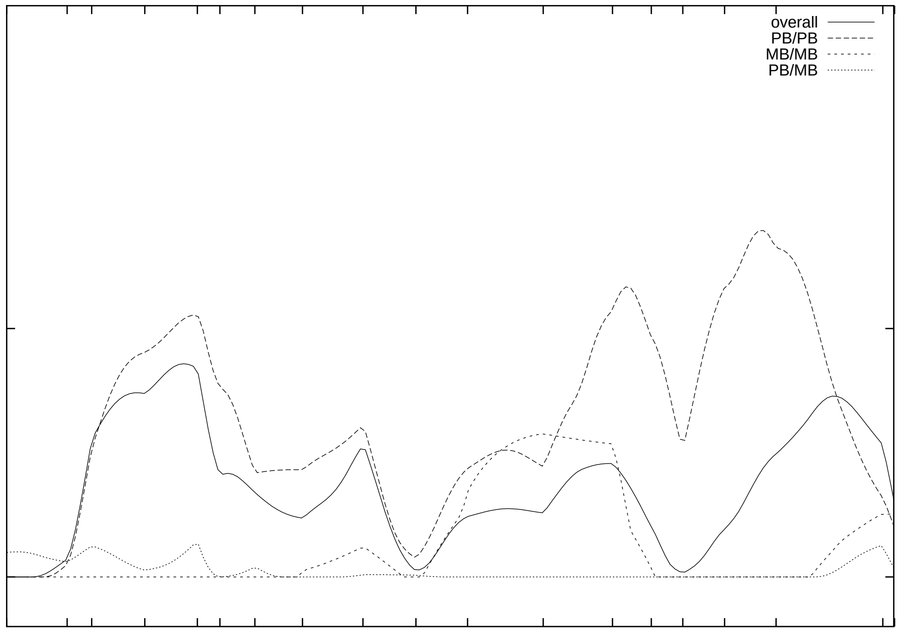


Figure 6.9: MLS curves for linkage analysis of affected sibships: chromosome 10. Results of the Siddiqui et al. (2001) scan of the same region are also shown.

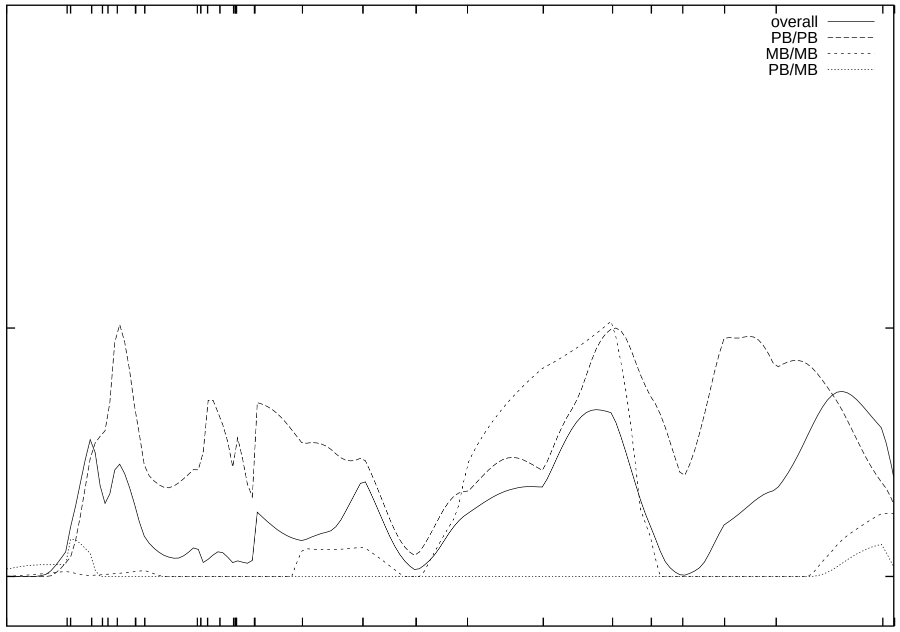
Chromosome 21 followup

In the second stage analysis for chromosome 21, 14 additional markers were typed around and beyond D21S266 in the nuclear families, and all 19 were typed in the members of the extended families identified for typing as described in section 6.4.1. After typing the additional markers in the nuclear families only, the evidence for linkage appeared to be reinforced. As shown in figure 6.11, the addition of these markers led to an increased multipoint MLS peak score of 1.1 to the right of the region first typed.

To analyse the data from the extended pedigrees and nuclear families together, allele sharing lod scores were calculated using the S_{all} and S_{pairs} scoring functions under the exponential model of Kong and Cox (1997b). The multipoint scores are shown in figure 6.12 and the single point scores in table 6.4. These results make the evidence for linkage somewhat less clear. There are single point scores above 1 in the 21q22.2 region (D21S1255, D21S1893), but these are in a neighbouring region to the multipoint peak in the nuclear families alone. The multipoint results show very little evidence for linkage, with a maximum lod of 0.7.



(a) Stage I markers



(b) Stage II markers

Figure 6.10: MLS scores for chromsome 10 by leprosy type

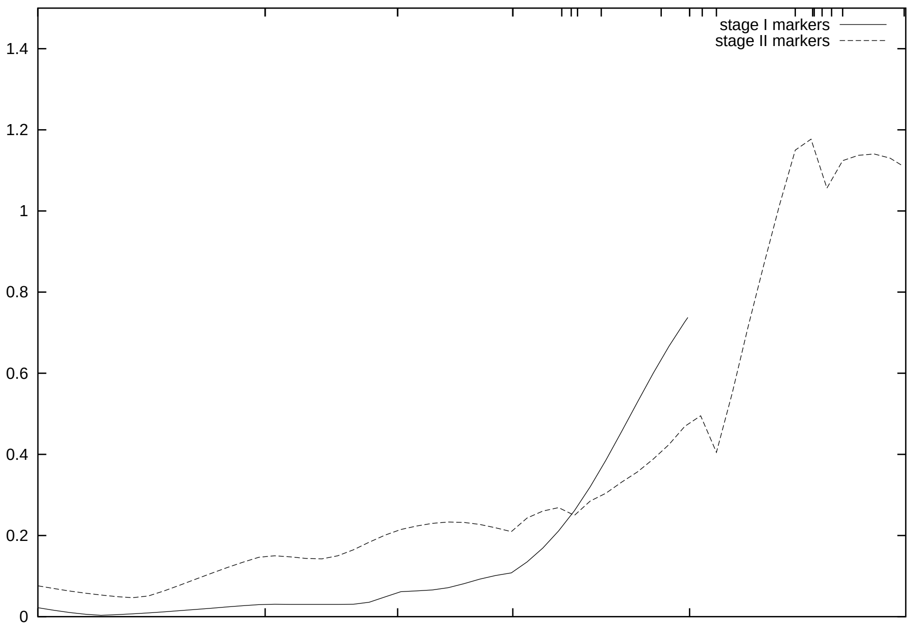


Figure 6.11: MLS scores among nuclear families on chromosome 21. The marks on the lower axis indicate the position of the stage I markers and those on the upper axis the position of the stage II markers

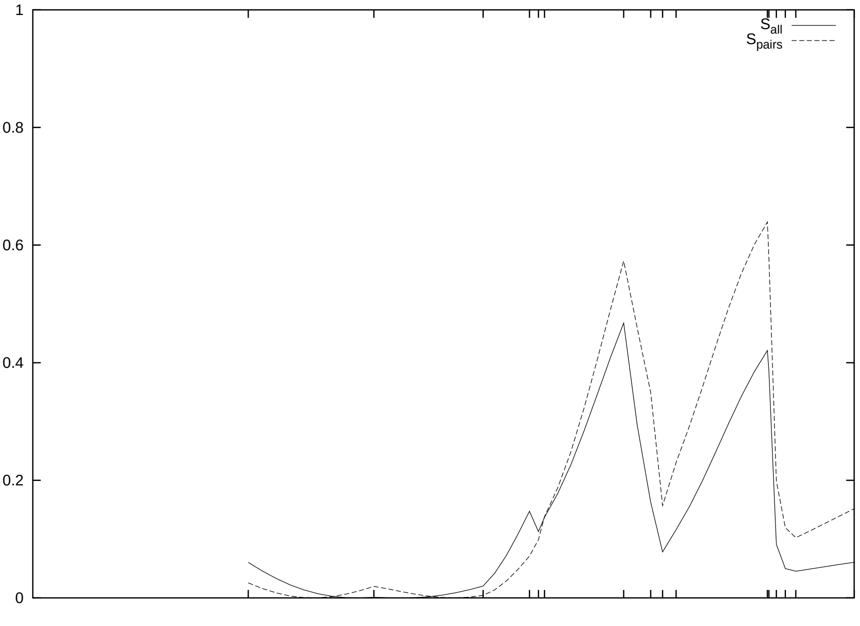


Figure 6.12: Allele-sharing lod scores from linkage analysis of chromosome 21 using extended pedigrees

Marker	location (cM)	lod score	
		S_{all}	S_{pairs}
D21S1914	23.0	0.0469	0.0136
D21S263	31.4	0.2387	0.1675
D21S1252	38.7	0.0647	0.0355
D21S267	41.8	0.6770	0.5342
D21S1891	42.4	0.3113	0.4137
D21S1255	42.8	1.3750	1.6429
D21S1893	48.1	1.1179	0.7878
D21S266	49.9	0.5724	0.8876
D21S1906	50.7	0.0531	0.0206
D21S1260	51.6	0.1151	0.2368
D21S1890	57.7	0.2491	0.4211
D21S1885	57.8	0.0009	0.0348
D21S1912	58.3	0.1617	0.2414
D21S1903	58.9	0.0295	0.0909
D21S1897	59.6	0.0513	0.0533
D21S2057	63.5	1.3206	2.0644

Table 6.4: Single point NPL scores for chromosome 21 markers used in the extended pedigree scan

6.4.3 Exclusion mapping

Exclusion maps are shown in figure 6.13. From the regions typed, 34.49%, 11.50% and 1.38% can be excluded from containing a gene with λ_S of 1.4, 1.6 or 2.0 respectively. The percent excluded are shown for each region in table 6.5.

6.4.4 Linkage to leprosy type

Subtype analysis was conducted by dividing the sibpairs (MLS analysis) or sibships (MLB analysis) into MB concordant, PB concordant and PB/MB mixed groups and analysing each group in turn. The numbers in each group are shown in table 6.6. The results of this subtype analysis among sibships (using all marker data including the additional chromosome 10 and 21 markers and described above), are shown in figure 6.14.

In this subtype analysis, there are peaks above 1 on chromosome 10 among PB concordant pairs and on chromosome X among discordant pairs. The score also approaches 1 on chromosome 15 and at the end of chromosome

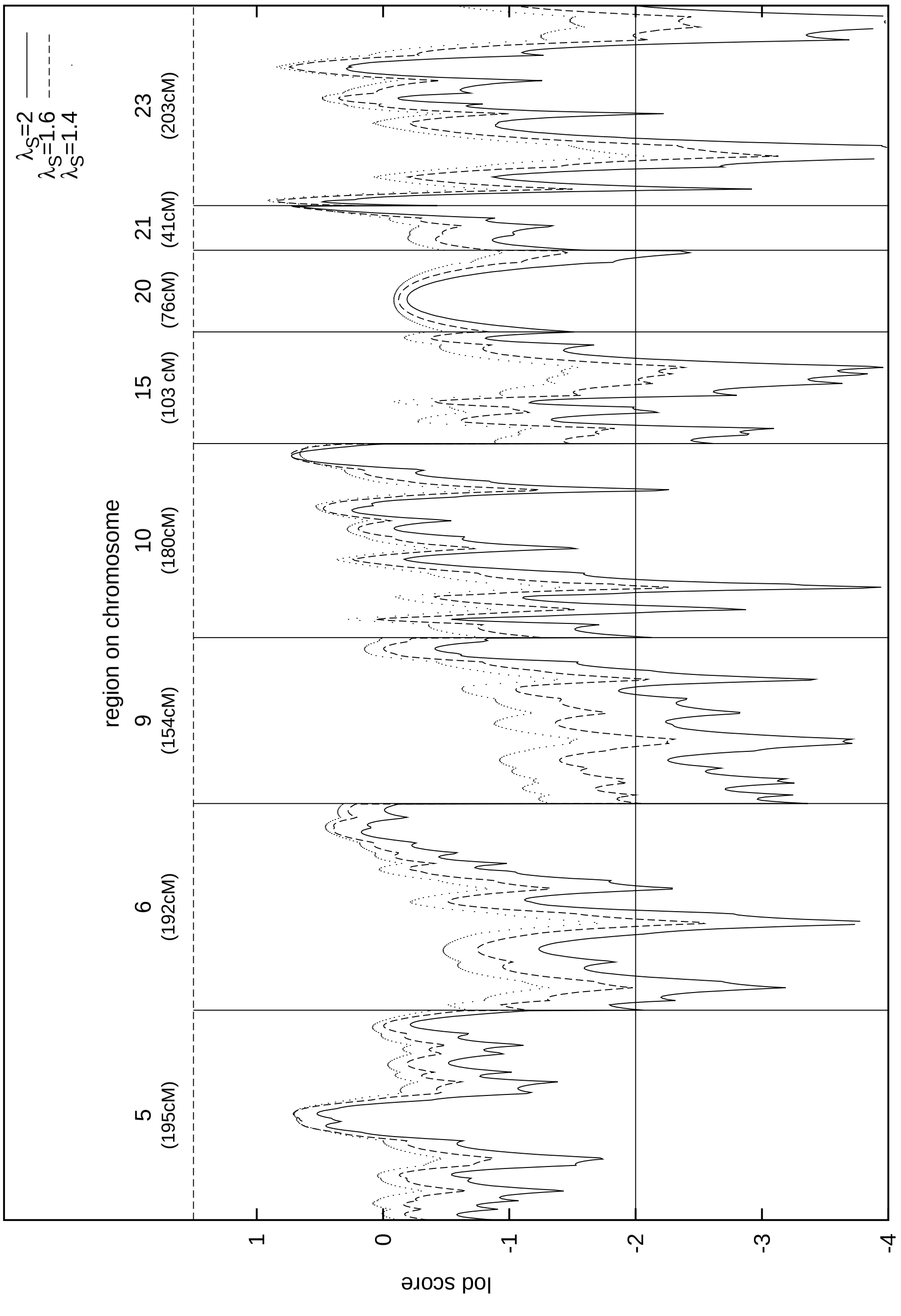
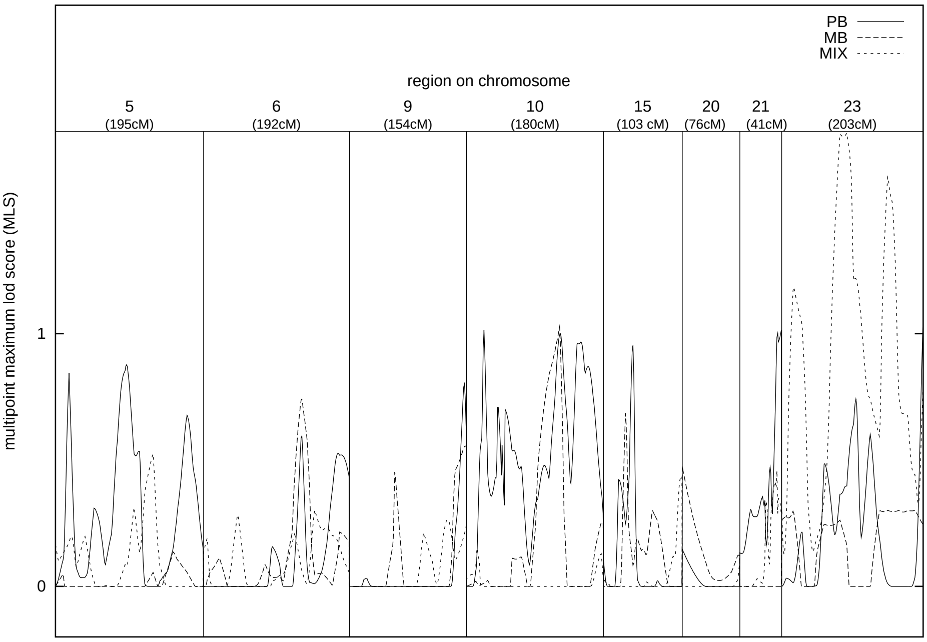
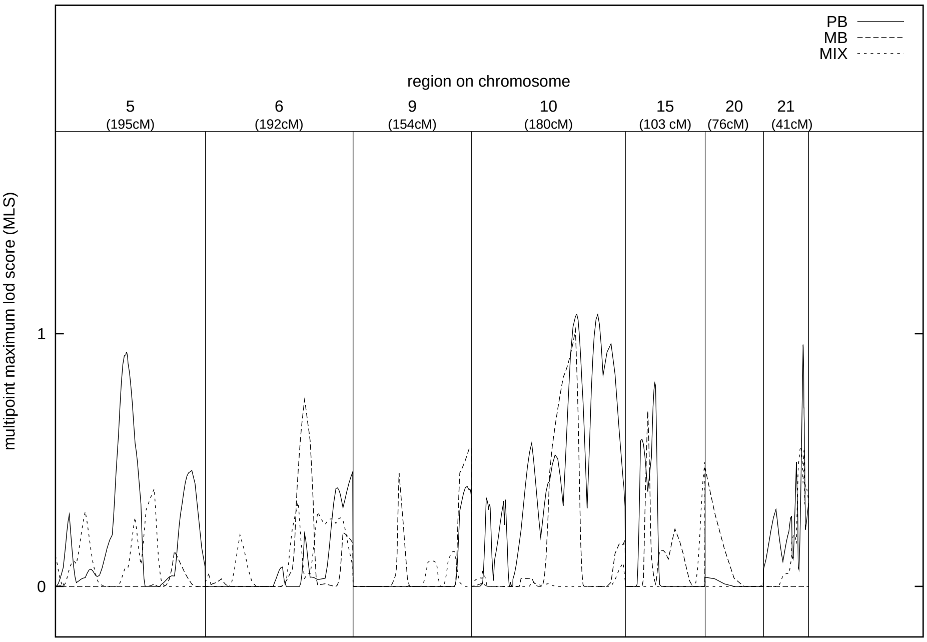


Figure 6.13: Results of exclusion mapping. The lines show the lod scores for $\lambda_S = 2, 1.6, 1.4$. A lod score less than -2 indicates the region can be excluded from containing a gene responsible for a genetic effect of at least λ_S.



(a) MLS curves



(b) MLB curves

Figure 6.14: Initial results from stage I: linkage analysis of leprosy among affected sibships, by leprosy type. Note: the MLB method is not suitable for the X chromosome

Chromosome	Length of region (cM)	Percent excluded for $\lambda_s =$		
		2.0	1.6	1.4
5	206	0.00	0.00	0.00
6	191	28.80	3.66	0.00
9	150	72.00	18.67	0.00
10	172	2.77	0.00	0.00
15	114	50.88	16.67	0.00
21	32	0.00	0.00	0.00
X	289	68.51	30.45	5.88
Total	1175	34.49	11.50	1.38

Table 6.5: Percent of chromosome regions excluded from nonparametric exclusion mapping

21, also among PB concordant pairs.

The highest MLS was on chromosome X, within discordant sibpairs. In one of these regions, there is also evidence for linkage among concordant pairs, as seen by the overall MLS score (figure 6.8). If a gene did play a role in determining the type of clinical leprosy which develops, one would expect excess IBD sharing among type concordant pairs and a deficit among discordant pairs. It is likely, then, that these high MLS scores reflect genetic influence over susceptibility to leprosy per se or simply random variation in this relative small sample of 26 type discordant pairs rather than any underlying genetic effect which determined clinical type of disease.

The results of the analysis of linkage to leprosy type among nuclear families using the method proposed by Holmans (2002) are shown in figure 6.15. The p values are very similar whether the two MB/MB pairs are included or excluded. Therefore, to keep power at a maximum, pairwise trait values were grouped into type-concordant and type-discordant.

Two regions, on chromosomes 10 and 15, have unadjusted p values below 1%. The p value is also low (~ 0.02) towards the end of chromosome 21, and this is interesting because it coincides with the region where there was confusing evidence for linkage to leprosy per se, as described above. The mean posterior IBD sharing probabilities in these regions are shown in figure 6.16 6.16. In the chromosome 10 and 15 regions, there are an excess of type concordant pairs sharing two alleles IBD, and an excess of type-discordant pairs sharing no alleles IBD. In the chromosome 21 region, however, there

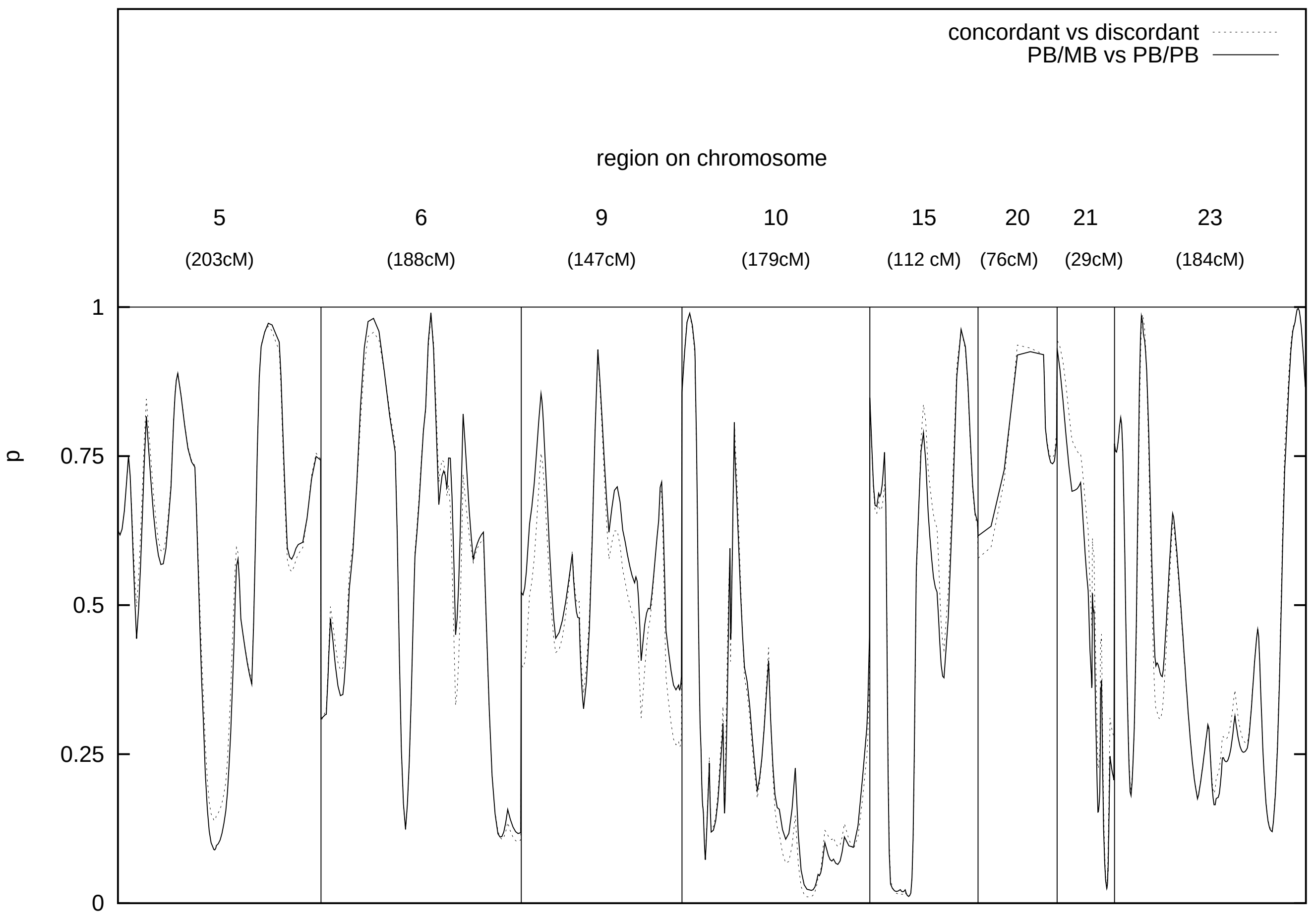
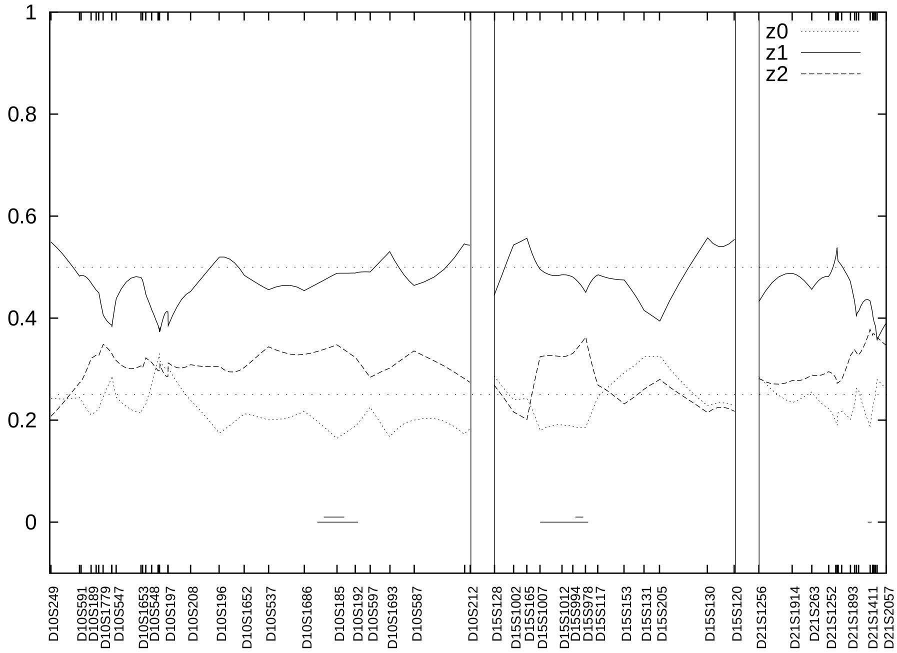
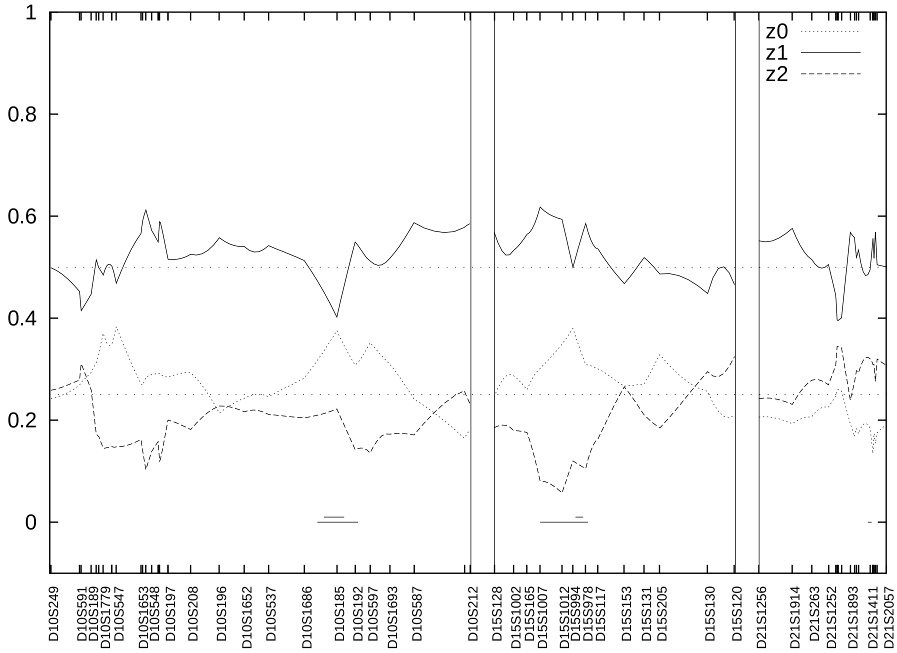


Figure 6.15: Results of analysis of linkage to leprosy type. A low p value is evidence for linkage



(a) Type concordant pairs



(b) Type discordant pairs

Figure 6.16: Posterior IBD sharing probabilities for chromosome 10, 15 and 21 regions. Regions with an unadjusted $p < 0.05$ *are marked with a solid horizontal line and with a double line where* $p < 0.01$*. Dotted lines are drawn at the expected IBD sharing probabilities of 0.25 and 0.5*

	Pairwise leprosy type			Total
	MB/MB	PB/PB	MB/PB	
# (non-independent) pairs	2	72	26	100
# sibships	2	56	24	82

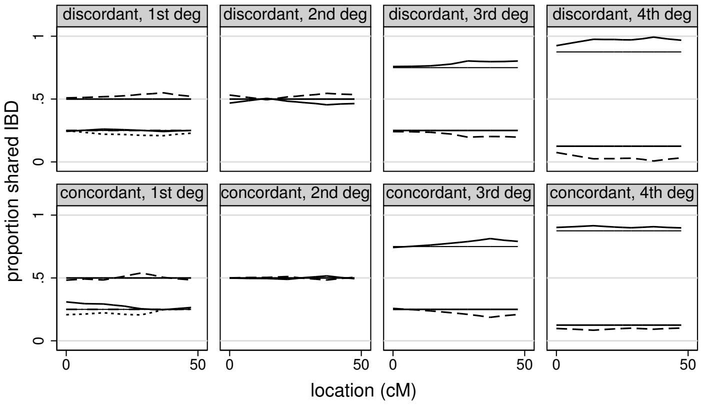
Table 6.6: Numbers of sibpairs and sibships by leprosy type concordance

is an excess of both type concordant and type-discordant pairs sharing two alleles IBD, though the proportions deviate from the expected values less for discordant than concordant pairs. The minimum unadjusted p values in each region are 0.006 (chromosome 10), 0.009 (chromosome 15) and 0.018 (chromosome 21) which do not reach genome wide significance levels, but were low enough to be considered worthy of further investigation.

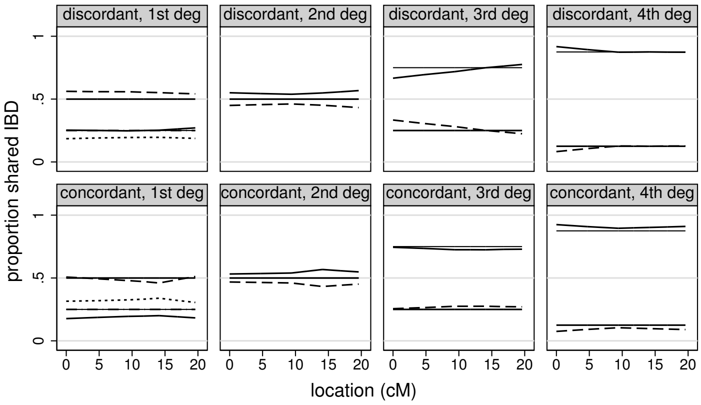
Ten markers in these regions where the p value was below 0.01 (6 on chromosome 10 and 4 on chromosome 15) were typed in the extended pedigree members (chromosome 21 markers had already been typed in the extended pedigrees). The number of affected relative pairs in the combined dataset of extended pedigrees and nuclear families (not including parent-offspring) is shown by degree and pairwise type in table 6.7. There are only eight MB/MB pairs, and, given the results from the analysis of nuclear families alone, it was decided to group these pairs into type concordant and type discordant.

The (multipoint) mean posterior IBD sharing, by concordance and degree of relationship, is shown in figure 6.17. Again, if a locus affected susceptibility to leprosy type, we would expect excess sharing among concordant pairs and a deficit among discordant pairs. For chromosome 10, this pattern is evident, if at all, only among 1st degree relatives (sibpairs). For chromosome 15, there is more obvious evidence of this kind of sharing among 1st degree relatives, but such a pattern is not clear among higher degree pairs. Towards the end of chromosome 21, however, we find excess IBD sharing among concordant pairs of all degree while among discordant pairs mean posterior IBD sharing is either at or below that expected under the null.

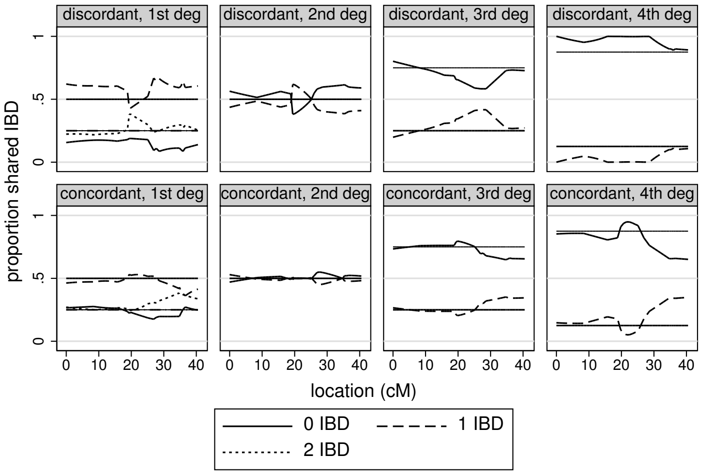
The results of the formal analysis are shown in figure 6.18. The evidence for linkage on chromosome 10 has decreased (minimum multipoint $p = 0.03$) but increased very slightly on chromosome 15 (minimum multipoint $p =$ 007). On chromosome 21, the minimum multipoint/singlepoint p values are



(a) Chromosome 10



(b) Chromosome 15



(c) Chromosome 21

Figure 6.17: Mean posterior IBD sharing among affected relative pairs, by pairwise concordance and degree of relationship

	Pairwise leprosy type			
Degree	MB/MB	MB/PB	PB/PB	Total
1	2	47	130	179
2	6	52	124	182
3	0	24	105	129
4	0	10	15	25
Total	8	133	374	515

Table 6.7: Number of affected relative pairs from extended families, by pairwise leprosy type and degree of relation

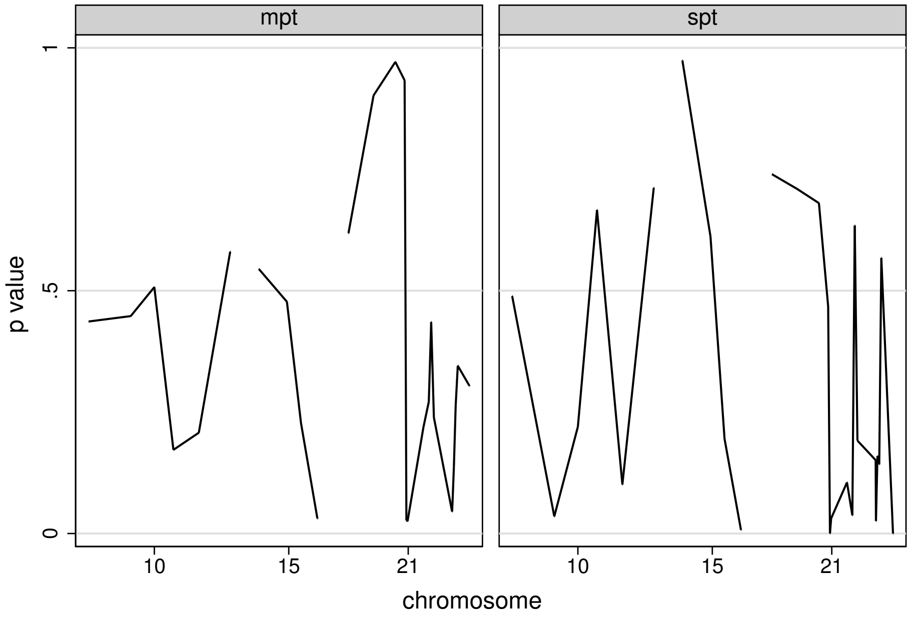


Figure 6.18: Linkage analysis of leprosy type on chromosomes 10, 15 and 21 among extended pedigrees: concordant vs discordant

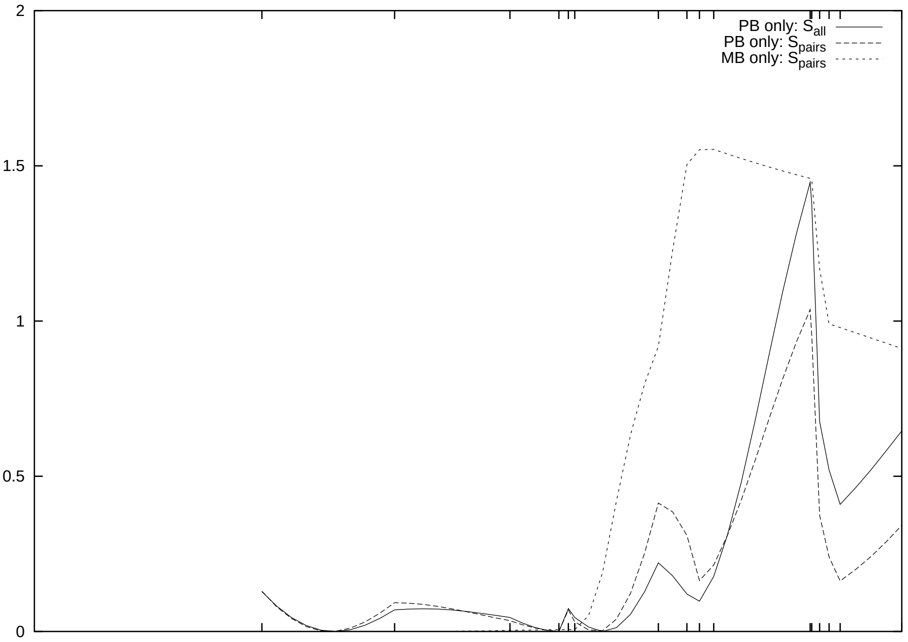


Figure 6.19: Multipoint allele-sharing lod scores for chromosome 21: considering PB and MB cases separately. Note there is no curve for S_{all} for MB cases because no pedigree contained more than two MB members and so S_{all} and S_{pairs} are equivalent

0.045/0.0001 (D21S2057).

We can also conduct a subtype analysis on these extended families by considering just PB cases to be affected and then just MB cases to be affected. Lod scores calculated using the NPL scores under the exponential model for such an analysis on chromosome 21 are shown in figure 6.19. Comparison with figure 6.12 shows that the maximum lod scores when PB and MB cases are considered separately are considerably higher (1.4 and 1.5 respectively) than when all cases are considered together (0.6).

6.5 Discussion

6.5.1 Selection of unaffected members from extended pedigrees

Holmans and Clayton (1995) and Holmans and Craddock (1997) have shown that the most efficient strategy in the first stage of ASP screens is generally to type additional affecteds rather unaffecteds. In this project, we could not recruit any additional affected individuals and needed to maximise power in the second stage of a genome screen which involved extended pedigrees. This was done by selecting those unaffected pedigree members who provided maximum expected information about IBD sharing among the affected members.

The analysis of template pedigrees led to simple rules for selection of unaffected members, as given in table 6.3. It is not so simple to find general rules for the more complicated extended pedigrees, but consideration of the results from the simulations led to some recommendations. Any unaffecteds who directly 'link' affected pedigree members should be typed if available, as should any common ancestors of the affected members. If both parents of an affected individual are available, they should be typed. If one or both are not available, and there are no affected siblings, one or more unaffected siblings will be useful to help reconstruct the parents' genotypes. This becomes less important if grandparents are included (because they are the common ancestors of this and other affecteds in the pedigree). In the case where there are no parents or unaffected siblings available, half siblings, then aunts or uncles may be useful. But they should be considered as a last option. More distant unaffected relatives are not generally useful.

The 'best subsets' chosen using simulation are broadly similar with those that might be expected to have been chosen without their use. The simulations were useful mainly for the very large pedigrees which contained only a few unaffected members, when the temptation may have been to type unnecessary members. In this project, a major motivation was to minimise genotyping because of cost, whilst maintaining power, and the use of simulation means this target has been met. However, given the time taken to run the simulations, if time and cost were equal priorities, selecting unaffected members using reasoning alone would be an equally, if not more suitable method.

6.5.2 Linkage analysis of leprosy per se and leprosy type

The analysis of nuclear families in stage I of the partial genome screen for linkage to leprosy per se and leprosy type showed three and two regions suggestive of linkage respectively.

Chromosome X

The chromosome X region has yet to be typed in the extended pedigrees and so only speculative discussion about this region is available at this point. It is likely that the high MLS score at the end of this region in the analysis of linkage to leprosy per se is an artifact, but the score of 1.2 in the DXS993–DXS991 region does appear to be worth further consideration. The MLS score is even higher ($\sim$ 1.8) among type discordant pairs in this region, indicating that if this is the result of a real underlying genetic effect, it would control susceptibility to leprosy per se and not leprosy type.

This peak is not, however, close to that found by Bellamy et al. (2000) which led to this region being studied (they are $\sim$ 100CM apart). Still, it is interesting and worthy of follow-up.

Chromosome 21

In the analysis of nuclear family data, MLS scores were above 1 among nuclear families after the additional markers were typed, but the allele sharing lod scores when extended pedigree data was also included were below 0.7 across the region. Among these extended pedigrees, there were two neighbouring markers with singlepoint lod scores above 1 (D21S1891, D21S1255), but these did not coincide with the peaks among nuclear families which were further towards the telomere (D21S1411).

When the nuclear family data was analysed separately by leprosy type, the MLS and MLB scores were around 1, again in the region of D21S1411 and when the extended pedigree data was also included, allele sharing lod scores were again above 1 in this region (1.4 for PB cases only, 1.5 for MB). The IBD regression analysis also indicated that there was some evidence for linkage to leprosy type (multipoint $p = 0.045$, singlepoint $p = 0.026$).

Overall, there is sufficient evidence to deem this region worthy of further followup among the Karonga population.

Chromosome 10

Initial results from chromosome 10 appeared to agree with the peak at 10p13 identified by Siddiqui et al. (2001). When extra markers were typed in this region, however, our peak disappeared (see figure 6.9). In the Indian study, all but two cases were PB and in the recent Vietnamese scan, evidence for linkage to this region was found only among PB concordant pairs. When Karonga sibs were analysed separately by leprosy type, the MLS for PB concordant pairs was above 1 in the neighbouring 10p14 region.

Since the partial genome screen in this project was initiated a strong positional candidate has been identified on chromosome 10p13 in the Indian population studied by (Siddiqui et al., 2001). The gene MRC1 encodes the macrophage mannose receptor C type 1 and a specific polymorphism in this gene has been identified to be associated with leprosy susceptibility in India (Tosh et al, in preparation). The same polymorphism has been investigated for association in a Malawi leprosy case-control study, however no evidence of association was found (Fitness et al., in preparation). The allele associated with leprosy susceptibility in Southern India occurs less frequently in Northern Malawi (7% vs 44%) and this may explain our failure to replicate Siddiqui's results, but it does not explain why we find suggestive evidence for linkage in a neighbouring region. Note that this region did not show strong evidence for linkage in IBD regression analysis of linkage to leprosy type ($p > 0.1$).

Chromosome 6

Perhaps the most surprising of these results is the lack of evidence for linkage in the HLA-DR region (chromosome 6p21). Although there has been no published association between HLA-DR2 and leprosy in the Malawian population, but it was expected that such an association would exist, given the number of other populations in which it has been found (see table 2.7.5). Our study did not exclude linkage in this region, but the exclusion mapping LOD scores were negative: -1.8, -1.0 and -0.6 for $\lambda_S = 2$, 1.6 and 1.4 respectively. These negative scores indicate evidence *against* linkage, but we lacked sufficient power to conclusively exclude this region at the standard LOD< -2 level. An association between HLA-DR and leprosy has gained general acceptance, and our negative finding is interesting, especially as diagnostic confirmation was so rigorous. Nor did we find evidence to support a locus in the 6q25–q27 region responsible for a locus-specific λ_S of over 2 (exclusion mapping lod score -0.57 for $\lambda_S = 2$).

6.5.3 Power

One use of exclusion mapping is to determine the power in a study. If positive linkage to a region cannot be determined, but neither can that region be excluded, the study probably lacks power. In our study, relatively few regions could be excluded using the data from the sibling pairs only. Leprosy is a complex disease with a relatively low overall genetic λ_S in the Karonga population of the order of 2 (estimated in chapter 4). Even if only 2 loci were responsible for increasing genetic susceptibility to leprosy, their locus-specific λ_S would be in the region of 1.4 if they acted multiplicatively to produce an overall λ_S of 2. Stage I of our study clearly does not have the power to detect even that effect.

This is to be expected as there were relatively few (82) sibships available. If more loci are involved (which is likely), then we can expect any locus-specific λ_S to be even lower. Any linkage study into susceptibility to leprosy must then be much bigger if it is to detect significant linkage. Indeed, Siddiqui et al. (2001) used 224 families (245 independent sib pairs) in their study in which they found significant linkage of susceptibility to leprosy per se to chromosome 10p13, with an estimated locus-specific λ_S of 1.6.

Power should be increased in stage II of this study, with the addition of extended pedigrees, but estimating power for the extended pedigrees is not easy. We could use simulation, but the design is quite complicated (stage I: type nuclear families; stage II: type extended pedigrees with closely spaced markers in regions where the MLS from stage I exceeds 1.0). Also, a mode of inheritance would have to be assumed, but leprosy does not have a known mode of inheritance.

6.6 Summary

This chapter describes a two-stage partial genome screen scan for linkage to leprosy per se and leprosy type among families from Karonga. In the first stage, 82 nuclear families were typed across eight regions on chromosomes 5, 6, 9, 10, 15, 20, 21 and X which had been identified in studies of genetic susceptibility to leprosy or TB in other populations. Three regions were identified as showing suggestive linkage to leprosy per se and two to leprosy type. In relation to linkage to leprosy per se, followup of regions of chromosomes 10 and 21 did not lead to increased evidence for linkage. The chromosome X region has yet to be followed up.

In relation to linkage to leprosy type, followup of regions on chromosomes 10 and 15 in extended families found the overall evidence for linkage decreased on chromosome 10 and increased slightly on chromosome 15. Further investigation of IBD sharing in these regions showed IBD sharing varied with leprosy type in siblings only. There was more consistent evidence for linkage to chromosome 21q22.3, with allele sharing lod scores approaching 1.5 when PB or MB cases were analysed separately and excess IBD sharing among type concordant pairs of up to fourth degree, confirmed by a p value of 0.045 in the IBD regression analysis. However, neither this nor any other result came close to the level of statistical significance required for genome-wide analysis.

CHAPTER 7.
DISCUSSION

Human immune response to infection by *M. leprae* varies considerably between individuals. Infection is necessary for, but does not always lead to disease which itself may be manifested across a spectrum from PB to MB. This spectrum of clinical disease is due to very different immune responses, the determinants of which are still not clear. In addition to genetic makeup, non-genetic factors such as past exposure to other mycobacteria influence immune response. Studies in several different populations have found evidence that at least one gene in the MHC class II region and other, apparently population-specific genes or chromosome regions are linked to or associated with susceptibility to leprosy or to clinical type.

The aims of this project were defined in section 1.1. This chapter summarises the work accomplished to meet those aims and discusses the potential for wider application of the methods developed. The evidence for genetic susceptibility to leprosy is summarised and, lastly, recommendations are made for future studies in this area.

7.1 Review of results relating to aims and objectives

7.1.1 Estimation of λ_R

λ_R is an important measure of genetic effect, and, as shown by Risch (1990b), can be used to predict the power to detect linkage using relative pairs. Familial clustering of disease may be due to shared non-genetic factors as well as shared genetic factors. Failure to account for non-genetic factors can lead to biased estimates of λ_R, which could in turn lead to over-optimistic expectations for the potential of linkage analysis studies. Thus, calculation of unbiased estimates of this parameter would be a useful preface to any linkage analysis to ensure adequate power.

In chapter 4, one method to estimate the excess risk to relatives of cases after accounting for known non-genetic covariates was proposed. In contrast to other models used in the study of familial disease, the proposed model did not depend on the availability of incidence data, but made use of the present-state data which was available from the KPS.

Risk of leprosy depends on many non-genetic factors, such as age, sex, BCG vaccination, exposure to environmental mycobacteria, household contact and socio-economic status. Of these, household contact is most likely to cluster within families, but there is also a tendency for familial clustering of socio-economic status, environmental mycobacteria exposure and (to a lesser extent) BCG vaccination. Estimates of λ_R which do not account for the effect of these factors may be substantially inflated. We found $\lambda_S \simeq 3$ and $\lambda_S \simeq 2$ before and after accounting for known non-genetic risk factors (although this is likely to be an upper limit for reasons discussed in section 4.8.2). This demonstrates the considerable bias that can result when non-genetic factors are ignored.

The proposed method has potential to be applied to other diseases which may be subject to both genetic and non-genetic factors. Although we had the relative luxury of dealing with data from a complete population survey, the method is not dependent on this and the model can be fitted using appropriate weights and robust standard errors if the sampling strategy is well-defined.

7.1.2 Efficient use of extended pedigrees in linkage analysis

Although there has been considerable work on the aggregation of disease among relative pairs and sibships, there has been little on more general combinations of affected relatives. Within the pedigrees identified from the KPS database were many extended multicase pedigrees. The shape and size of these pedigrees varied considerably, from small nuclear families to one pedigree of 26 bits which contained 21 members, nine of whom were affected. It was likely that these larger pedigrees would provide greater power to detect linkage than the nuclear families alone.

Aggregation of disease in families and implications for design of linkage analyses

The work presented in chapter 4 explored aggregation of disease among relative trios by considering the behaviour of the 'second degree relative recurrence risk ratio' under selected single-locus models. It was shown that, under such models, affected relative trios allow for greater discrimination between genetic models than affected relative pairs. The dependence of λ_S on the genetic model was explored and it was shown that, while it may be convenient to treat λ_S as an independent parameter, the valid range of λ_S for the genetic model under consideration must be noted if results are to be interpreted sensibly.

Power to detect linkage under a specific model may be calculated using the same methods, and this was demonstrated for half siblings. The results showed that a strategy of recruiting half siblings is likely to be more powerful than full siblings in a linkage study unless the underlying genetic model is strongly recessive and that half sibling trios provide substantially more power than pairs. This result does not directly allow us to decide which affected members of a pedigree should be included in a linkage analysis. It does, however, add weight to the argument that larger sets of affected relatives will offer more power than relative pairs in linkage analysis.

In this project, because of the limited number of pedigrees available, it was decided that the best approach would be to include all pedigrees which contained at least two affected members.

Sampling strategies for unaffected pedigree members

Once it has been decided that particular affected individuals will be typed in a linkage study, additional unaffected relatives may also be typed for two reasons:

1. to confirm reported genetic relationships (eg typing parents of an affected sibling pair can be useful for determining mispaternities);
2. to help resolve IBD sharing among affected members.

The choice of unaffected members for the second reason was examined in chapter 6 using simulation. The expected information about IBD sharing among affected members in the absence of linkage was compared under different selection strategies for unaffected members. It was shown that when marker polymorphism is low and marker spacing wide, unaffected members can provide a substantial increase in information about IBD sharing. In this situation, the choice of a 'best subset' depends on the pedigree available and general recommendations were presented in section 6.5.1. As marker polymorphism increases and marker spacing decreases, IBS sharing approximates IBD sharing and so the typing of unaffected individuals becomes less important, although they still offer increased information in most cases.

With the marker spacing used in the first stage of most genome scans ($\sim$ 7–10cM), some unaffected relatives should generally be included in any study, partly to identify misreported genetic relationships, but also to increase power in studies such as this when the sample size is relatively small and ascertainment of additional families is not possible. However, the choice of which unaffected relatives to include can probably just as well be made on the basis of common sense. The method of simulation used here, although intuitively simple, did take a very long time to run.

The linkage analysis presented here employed a two-stage strategy. Initially, regions identified through a review of published studies in genetic susceptibility to leprosy were typed using a 7–10cM marker grid among 83 nuclear families (after exclusions due to mispaternities). These data were analysed in three ways:

1. nonparametric linkage analysis was used to identify regions of susceptibility to leprosy per se;
2. both subgroup analysis and the method of IBD regression developed by Holmans (2002) were used to examine linkage to leprosy type;
3. exclusion mapping was used
   (a) to explore which regions could be assumed *not* to contain a gene responsible for particular magnitudes of genetic influence given our data
   (b) to give an idea of power to detect linkage. If we could neither exclude nor confirm evidence for linkage in a region, then we lacked power to detect a genetic effect of a particular magnitude.

Regions which showed suggestive evidence of linkage in this first stage were then scheduled to be typed among the extended pedigrees in order to investigate whether further evidence for linkage could be found.

Five regions were identified which showed suggestive evidence of linkage to leprosy or clinical type in the first stage of the screen. These were discussed in section 6.5.2; briefly, two regions remain which we consider worthy of followup: chromosome Xp11, in relation to leprosy per se, and chromosome 21q22, in relation to leprosy type.

7.2 Evidence for genetic influence of human susceptibility to leprosy

7.2.1 Relating the results of this project to published work elsewhere

The evidence for genetic influence of human susceptibility to leprosy was reviewed in chapter 2. There is good evidence from previous studies that MHC class II genes are involved, and a positive association between both tuberculoid and lepromatous disease and the DR2 and DQ1 antigens and strong evidence for linkage to this region have been found in several populations. There is also contradictory evidence for association of the TNF2 allele of the MHC class III TNF-α gene with increased and decreased risk of leprosy in Central Indian and Brazilian populations respectively (Santos et al., 2000, 2002; Shaw et al., 2001; Roy et al., 1997).

Neither the recent genome screen in India (Siddiqui et al., 2001) nor the partial genome screen in this project found any evidence for linkage to the MHC, but reasons for this are unclear. Their study was adequately powered and association with HLA-DR has been repeatedly observed in Indian populations (although there are no published reports of association studies in the specific regions studied by Siddiqui et al. (2001)). Although linkage to this region was not excluded in the Karonga population at the standard LOD< -2 level, negative exclusion mapping LOD scores provided evidence against linkage in this region.

Recent studies found significant evidence for linkage on chromosomes 10p13 and 20p12 in India (Siddiqui et al., 2001; Tosh et al., 2002), near HLA/TNF in Brazil (Shaw et al., 2001) and 6q25 in Vietnam (Mira et al., 2003). None of these regions showed evidence for linkage in the Karonga families analysed in this project and the allele subsequently identified in the Indian study as associated with leprosy is at a much lower frequency among the Karonga population (7% vs 50%).

7.2.2 Explanations for inconsistency between studies

A recent review of association and linkage studies of leprosy by Fitness et al. (2002) referenced nearly 180 articles and concluded:

> Many of the associations have been found in a small series of patients, or in a single population and should be repeated in larger studies. Lack of correlation in results between populations should not necessarily be regarded as a negation of initial associations; but may instead reflect heterogeneity in the genetic susceptibility to this enigmatic disease.

Under the theory of genetic heterogeneity, several different genes may affect susceptibility to disease. If the distribution of alleles present at these loci differ between populations, as is likely to be the case, association or linkage studies in different populations will find different results.

Significant results for *NRAMP1* in two South-East Asian studies were not replicated in studies of Brazilian, French Polynesian, Indian or Pakistani populations. Despite well-conducted studies finding evidence for association with VDR (India) and Laminin $\alpha 2$ (Indonesia), there are no published replications of these findings.

Many infectious disease agents (eg, *Plasmodium falciparum* and *M. tuberculosis*) will have a strong influence on selection of genes associated with the immune response. The ecology of these infections can differ greatly between populations and it is thus expected that there will be differences between populations in the many genes which determine susceptibility to infectious disease.

Leprosy is likely to provide less significant selective pressure than TB, for example, because it is generally a less severe disease. On the other hand, it has been suggested that the same genes might control susceptibility to both TB and leprosy in humans. There are obvious biological similarities between the two diseases - both are caused by infection with mycobacteria and only a small proportion of those infected by either *M. leprae* or *M. tuberculosis* develop disease. In mice, the same gene (*Nramp1*) controls susceptibility to both diseases. TB is an increasing public health problem worldwide (Dye et al., 1999) and there have been many studies of genetic susceptibility to TB (see Abel and Casanova (2000) for a review).

Recent studies have suggested that there is heterogeneity between populations in the genetic control of TB. For example, Delgado et al. (2002) studied genetic variants found to be associated with susceptibility or resistance to TB in other ethnic groups among 358 TB patients and 106 tuberculin-positive controls in Cambodia. They failed to replicate many of the findings from other populations, but also found evidence that variants associated with susceptibility in at least two other populations were significantly associated with resistance in Cambodia (see Delgado et al., 2002, table 3). If the same genes do control immune response to infection by *M. leprae* and *M. tuberculosis* and if there is genetic heterogeneity in susceptibility to TB, then it is likely that there is also genetic heterogeneity in susceptibility to leprosy.

Genetic differences between populations due to the selective pressure of TB and other diseases or because of genetic drift may then be a reasonable explanation for the variety of results seen in different populations for genetic susceptibility to leprosy. Similarly, if *M. leprae* itself differs between geographical regions, this could lead to apparent differences between genetic analyses of the human population in those regions.

The TNF2 allele been reported to be significantly associated with both increased (in Central India) and decreased (in Brazil) risk of disease. However, genetic heterogeneity alone would not explain the contradictory findings for TNF2. If it was not TNF but another gene which affected susceptibility to leprosy, and if this other gene was in linkage disequilibrium with TNF in such a way that the high risk allele was found with TNF2 in Central India and with TNF1 in Brazil, this could lead to such observations.

Another possible explanation for the variation in published findings is that some positive results are attributable to multiple testing. When n independent statistical tests are performed, each with a significance level of α, one of them will be positive under the null with probability $1-(1-\alpha)^n$. For this reason, when multiple tests are performed, p values should be adjusted to account for the number of tests performed.

This would explain why evidence for the involvement of some genes or chromosome regions in addition to MHC has been reported in studies from various different populations, but rarely confirmed in other studies. It would also explain some of the reported associations with HLA antigens which have not been replicated in other studies, although they might also be due to linkage disequilibrium with genes which are associated with leprosy susceptibility. This is an alternative explanation for the conflicting results from the TNFα studies - the TNF2 allele has been found to be in linkage disequilibrium with HLA-DR alleles (Roy et al., 1997).

Gene-environment interaction provides another explanation for the variation in results to date, in so far as genetic susceptibility may be expressed only in the presence of particular non-genetic factors. To take a simple example, suppose people are either genetically ‘responsive’ or ‘non-responsive’ to BCG vaccination. Those people who are non-responders would appear to be more susceptible to leprosy only in a population where most people were vaccinated. Thus, a study carried out in a population which had a high rate of vaccination might find evidence for a genetic effect while a similar study in a population with low vaccination rates would not. Note this difference does not depend on genetic heterogeneity but on environmental heterogeneity.

All of the above are possible explanations for the differences in published results and these differences are likely to be due to a combination of false positives, genetic heterogeneity and environmental heterogeneity in the presence of gene-environment interaction. The evidence in support of the genetic heterogeneity argument is strengthened by the lack of evidence for linkage to HLA in the study by Siddiqui et al. (2001), despite repeatedly published associations of leprosy and HLA-DR in other Indian populations. The same study found evidence for linkage to a region on chromosome 20p12 only in a geographically defined subset of families. Recall also that strong evidence for both positive and negative associations of TNF2 with leprosy susceptibility have been found in India and Brazil, respectively.

7.2.3 Genetic susceptibility to leprosy in the Karonga population

Fewer studies of leprosy have been conducted in Africa than either India or Brazil, but given the results of this project, it appears that genetic influence on leprosy susceptibility is different in Malawi when compared to either India or Brazil. We estimated that the sibling recurrence risk for leprosy among the Karonga population was ~ 2 after controlling for measured non-genetic risk factors. This is likely to be an upper limit, and taken together with the overall lack of evidence from our (underpowered) linkage analysis, it would be fairly straightforward to construct an argument *against* genetic influence of susceptibility to leprosy among the Karonga population.

This contrasts with the evidence from other populations summarised above and with the magnitude of locus-specific effects found by Siddiqui et al. (2001), Shaw et al. (2001) and Mira et al. (2003) - $\lambda_S \sim 1.66$ for 10p13 in a Central Indian population, $\lambda_S \sim 1.79$ for HLA in Brazil and $\lambda_S \sim 2.2$ for a region on chromosome 6q25 respectively. Although leprosy is now less prevalent in Karonga than either Brazil or India, incidence was higher in Karonga than Brazil as recently as the 1980s, so the current low incidence in Karonga cannot explain the lack of positive findings.

With several genes likely to play a role in this complex disease, and an upper limit of $\lambda_S \simeq 2$, it seems very unlikely a single gene in Malawi will be responsible for a locus-specific λ_S as large as those found in the Indian or Brazilian studies. This may be because the specific alleles which influence susceptibility to leprosy in India and Brazil are at much lower frequencies in Karonga - this certainly appears to be the case with the MRC1 allele implicated in Southern India.

7.3 Future work

7.3.1 Genetic studies of leprosy

The effect of human genes on leprosy susceptibility is likely to be heterogeneous. Positive results from studies in one population are often not replicated in another and it is not clear whether this is due to genetic or environmental heterogeneity between populations or false positive or negative results. For this reason, attempts to replicate positive results from genetic studies in one population should first be conducted in the same population. Once successfully replicated in the same population, failure to replicate in other populations can more reliably be determined to be due to genetic heterogeneity.

In particular, studies should be conducted which attempt to replicate association between leprosy and *NRAMP1* in South-East Asia; leprosy type and *NRAMP1* in Mali; *VDR* and leprosy in central India; *Laminin* $\alpha 2$ and leprosy type in Indonesia; linkage between leprosy and 20p12 in central India and between leprosy and 6q25 in Vietnam.

Two other results are particularly worthy of follow-up. First, it is not at all clear why Siddiqui et al. did not find evidence to support linkage to HLA in an Indian population. As a first step, association studies should be conducted in the same population Siddiqui et al. sampled to confirm whether or not this is due to a genuine lack of association or because the sample was too small to detect linkage in this region. Association studies of MHC genes in the Karonga population have found no significant results.

Second, though it is still controversial whether the TNF2 allele affects TNFα production, TNFα levels are raised among tuberculoid compared to lepromatous patients and three studies have found the TNF2 allele to be associated with leprosy. However, it is not clear why the direction of the association differs between Brazilian and Indian studies; further association studies are needed to examine this. Such studies should also examine neighbouring genes to TNFα to test whether the observed association is due to linkage disequilibrium between TNFα and some other gene. The strongest genetic association with TNF2 was observed with lepromatous leprosy, and accurate diagnosis of clinical type of leprosy will be necessary in any studies in this area.

Studies to specifically examine gene-environment interaction are particularly difficult for a disease such as leprosy because exposure to *M. leprae* cannot be accurately determined, incubation periods are measured in the order of years and immune response may be modified by lifetime exposure to other infectious agents. Therefore such studies should only proceed in populations in which evidence for genetic susceptibility to leprosy is stronger than it appears to be in Karonga.

Given the evidence for genetic heterogeneity, it would be interesting to compare estimates of λ_S from different populations, particularly the Indian, Vietnamese and Brazilian populations in which locus-specific recurrences risks have been estimated for regions identified in linkage analyses. Comparison of these locus-specific risks with the overall λ_S would indicate whether there are likely to be further genes of sufficient strength to be identified in linkage analysis studies.

7.3.2 Studies in the KPS

It is important to complete stage II of the linkage analysis begun in this project and to analyse the data from extended pedigrees in the follow-up of the regions that have shown potential linkage in stage I, in particular on chromosome X. If these analyses increase the evidence for linkage, DNA samples should be collected from additional leprosy affected families to confirm this result. These families would need to be identified from historical cases in the KPS database since leprosy has virtually disappeared from Karonga. Identification of all multiplex families would probably provide most information, but it would be simpler to target small families initially. All families containing affected (full) sibling pairs have already been targeted in Karonga and a strategy of targetting families through identification of affected half sibling pairs is recommended.

7.4 Summary

Several studies have reported significant results in relation to human genetic susceptibility to leprosy and there is fairly convincing evidence for at least the human MHC (see chapter 2). This thesis arose in order to apply linkage analysis methods to extended pedigree data from the KPS in an attempt to replicate some of these results in the Karonga population. Preliminary epidemiological analysis of the KPS data (chapter 3) showed that an individual's risk of disease depended on non-genetic factors such as age, BCG vaccination and household contact with infected individuals.

Marginal models were used to explore the aggregation of disease among relative pairs (chapter 4) and strong evidence was found for clustering of disease among families ($\hat{\lambda}_S \sim 5$). However, after adjusting for measured non-genetic factors, the estimated sibling recurrence risk, $\hat{\lambda}_S$, was under 2. This indicates that, in the Karonga population, susceptibility to leprosy is under greater influence from non-genetic factors which tend to aggregate in families such as household contact than genetic factors.

Calculation of the first and second degree recurrence risks under single locus genetic models (chapter 5) demonstrated that affected relative trios and above were likely to provide greater power than relative pairs in linkage analysis and a two-stage partial genome screen was conducted (chapter 6). The first stage analysed nuclear family data only while the second stage used extended pedigree data to followup regions of possible linkage found in stage I.

As might be expected from the low λ_S, no significant evidence for linkage was found, but two regions (Xp11 and 21q22) showed potential evidence for linkage (lod>1) and further followup of these regions is recommended.

However, the estimated overall $\hat{\lambda}_S$ is lower than some locus-specific λ_S found in other populations (eg Vietnam), which indicates likely heterogeneity between populations in respect of genetic susceptibility to leprosy.

Taken together, these results do not contradict those from other populations which indicate susceptibility to leprosy is influenced by host genetics. However, they suggest that the magnitude of genetic influence may be weaker in the Karonga population when compared with other populations. This is perhaps due to genetic heterogeneity resulting from differences in selective pressure from infectious agents in this compared to other populations.

BIBLIOGRAPHY

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

Abel L, Casanova JL (2000). Genetic predisposition to clinical tuberculosis: bridging the gap between simple and complex inheritance. Am J Hum Genet 67:274–277

Abel L, Demenais F (1988). Detection of major genes for susceptibility to leprosy and its subtypes in a Carribean island: Desirade Island. Am J Hum Genet 42:256–266

Abel L, Demenais F, Baule MS, Blanc M, Muller A, Raffoux C, Millan J, Bois E, Babron MC, Feingold N (1989). Genetic susceptibility to leprosy on a Caribbean island: linkage anaysis with five markers. Int J Lepr Other Mycobact Dis 57(2):465–471

Abel L, Demenais F, Preata A, Souza A, Dessein A (1991). Evidence for the segregation of a major gene in human susceptibility/resistance to infection by *Schistosoma mansoni* is associated with IgG reactivity to a 37-kDa larval surface antigen. Journal of Imunnology 140:2727–2736

Abel L, Müller-Myhsok B (1998). Robustness and power of the maximum-likelihood-binomial and maximum-likelihood score methods, in multipoint linkage analysis of affected sibship data. Am J Hum Genet 63:638–647

Abel L, Sánchez FA, Oberti J, Thue NV, Hoa LV, Lap VD, Skamene E, Lagrange PH, Schurr E (1998b). Susceptibility to leprosy is linked to the human *NRAMP1* gene. J Infect Dis 177:133–45

Abreu PC, Greenberg DA, Hodge SE (1999). Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. Am J Hum Genet 65(3):847–857

Alcaïs A, Sanchex F, Thus N, Lap V, Oberti J, Lagrange P, Schurr E, Abel L (2000). Granulomatous reaction to intradermal injection of lepromin (Mitsuda reaction) is linked to the human NRAMP1 gene in Vietnamese leprosy sibships. J Infect Dis 181:302–308

Bale U, Contractor N, Bhatia H (1985). HLA segregation study in families of leprosy patients. Indian J Med Res 82:198–201

Bellamy R, Beyers N, McAdam KP, Ruwende C, Gie R, Samaai P, Bester D, Meyer M, Corrah T, Collin M, Camidge DR, Wilkinson D, van Helden EH, Whittle HC, Amos W, van Helden P, Hill AV (2000). Genetic susceptibility to tuberculosis in Africans: a genome wide scan. Proc Natl Acad Sci USA 97(14):8005–8009

Bellamy RJ, Ruwende C, Corrah T, McAdam K, Whittle HC, Hill AV (1998). Variations in the *NRAMP1* gene and susceptibility to tuberculosis in West Africans. New Eng J Med 338(10):640–644

Besag J (1975). Statistical analysis of non-lattice data. The Statistician 24(3):179–195

Besag J (1997). Efficiency of pseudolikelihood estimation for simple Gaussian fields. Biometrika 64(3):616–618

Blackwelder W, Elston RC (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97

Brinkman B, Auijdeest D, Kaijzel E, Breedveld F, Verweij C (1995). Relevance of the tumor necrosis factor alpha (TNF alpha) -308 promoter polymorphism in TNF alpha gene regulation. J Inflamm 46(1):32–41

Bryceson A, Pfaltzgraff RE (1990). Leprosy. Churchill Livingstone

Bugawan T, Klitz W, Blair A, Erlich H (2000). High-resolution HLA class I typing in the CEPH families: analysis of linkage disequilibrium among HLA loci. Tissue Antigens 56(5):392–404

Buu N, Sánchez F, Schurr E (2000). The *Bcg* host-resistance gene. Clin Infect Dis 31(S3):S81–5

Carroll R, Ruppert D, Stefanski L (1995). Measurement error in nonlinear models. Number 63 in Monographs on statistics and apploied probability. London: Chapman and Hall

Cervino AC, Curnow RN (1997). Testing candidate genes that may affect susceptibility to leprosy. Int J Lepr Other Mycobact Dis 65(4):456–460

Chakravartti M, Vogel F (1973). A twin study on leprosy. In P. E. Becker, W. Lenz, F. Vogel, and G. G. Wendt (Eds.), *Topics in human genetics*, Volume 1, pp. 1–123. Georg Thieme Publishers

Chatterjee N, Shih J (2001). A bivariate cure-mixture approach for modeling familial association in diseases. Biometrics 57:779–786

Chatterjee N, Shih J, Hartge P, Brody L, Tucker M, Wacholder S (2001). Association and aggregation analysis using kin-cohort designs with applications to genotype and family history data from the Washington Ashkenazi study. Genet Epidemiol 21:123–138

Chilima BZ (2001). The natural history of the genus *Mycobaterium* in Karonga district, Northern Malawi. Ph. D. thesis, University of London

Chirwa TF (2001). Effect of household dynamics on risk of disease associated with household contact. Ph. D. thesis, University of London

Clayton D (2002). ibdreg

Clayton DG (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65(1):141–51

Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J (1986). Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Cooke GS, Hill AV (2001). Genetics of susceptibility to human infectious disease. Nat Genet 2:967–977

Cudworth G, Woodrow J (1975). Evidence for HLA-linked genes in 'juvenile' diabetes mellitus. Br Med J 3:133–135

Curtis D, Sham PC (1995). Model-free linkage analysis using likelihoods. Am J Hum Genet 57:703–716

Dale JR (1986). Global cross-ratio models for bivariate, discrete, ordered responses. Biometrics 42:909–917

Davis S, Weeks DE (1997). Comparison of nonparametric statistics for detection of linkage in nuclear families: Single-marker evaluation. Am J Hum Genet 61(6):1431–1444

de Vries R, Fat RLA, Nijenhuis L, van Rood J (1976). HLA-linked gentic control of host response to Mycobacterium leprae. Lancet 2:1328–1330

de Vries R, Mehra N, Vaidya M, Gupte M, Khan P, van Rood J (1980). HLA-linked control of susceptibility to tuberculoid leprosy and association with HLA-DR types. Tissue Antigens 16:294–304

Delgado JC, Naena A, Thim S, Goldfeld AE (2002). Ethnic-specific genetic associations with pulmonary tuberculosis. J Infect Dis 186:1463–1468

Dessoukey M, el Shiemy S, Sallam T (1996). HLA and leprosy: segregation and linkage study. Int J Dermatol 35(4):257–64

Duffy DL, Montgomery GW, Hall J, Mayne C, Healey SC, Brown J, Boomsma DI, Martin NG (2001). Human twinning is not linked to the region of chromosome 4 syntenic with the sheep twinning gene *FecB*. Am J Med Genet 100:182–186

Durner M, Vieland VJ, Greenberg DA (1999). Further evidence for the increased power of LOD scores compared with nonparametric methods. Am J Hum Genet 64:281–289

Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC (1999). Global burden of tuberculosis. Estimated incidence, prevalence and mortality by country. J Am Math Assoc 282:677–686

Feitosa M, Borecki I, Krieger H, Beiguelman B, Rao D (1995). The genetic epidemiology of leprosy in a Brazilian population. Am J Hum Genet 56:1179–1185

Fine PE (1982). Leprosy: the epidemiology of a slow bacterium. Epidemiologic Reviews 4:161–188

Fine PE (1988). Implications of genetics for the epidemiology and control of leprosy. Phil Trans R Soc Lond B 321:365–376

Fine PE, Pönnighaus JM, Maine N (1989). The distribution and implications of BCG scars in northern Malawi. Bulletin of the World Health Organization 67(1):35–42

Fine PE, Sterne J, Pönnighaus JM, Bliss L, Saul J, Chihana A, Munthali M, Warndorff D (1997). Household and dwelling contact as risk factors for leprosy in northernMalawi. Am J Epidemiol 146(1):91–102

Fine PE, Wolf E, Pritchard J, Watson B, Bradley D, Festenstein H, Chacko C (1979). HLA-Linked genes and leprosy. J Infect Dis 140(2):152–161

Fitness J, Floyd S, Warndorff DK, Sichali L, Mwaungulu L, Crampin A, Fine PE, Hill AV. Large scale candidate gene study of leprosy susceptibility in Karonga District, Northern Malawi. in preperation

Fitness J, Tosh K, Hill AV (2002). Genetics of susceptibility to leprosy. Genes and Immunity 3(8):441–453

Frank M (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. Aequationes Mathematicae 19:194–226

Gorodezky C, Flores J, Arevalo N, Castro L, Silva A, Rodriguez O (1987). Tuberculoid leprosy in Mexicans is associated with HLA-DR3. Lepr Rev 58:401–6

Greenberg DA, Abreu PC, Hodge SE (1998). The power to detect linkage in complex disease by means of simple LOD-score analyses. Am J Hum Genet 63:870–879

Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000). Allegro, a new computer program for multipoint linkage analysis [letter]. Nat Genet 25:12–13

Guo SW (1998). Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. Am J Hum Genet 63:252–258

Guo SW (2000). Familial aggregation of environmental risk factors and familial aggregation of disease. Am J Epidemiol 151(11):1121–1131

Haile R, Iselius L, Fine PE, Morton N (1985). Segregation and linkage analyses of 72 leprosy pedigrees. Hum Hered 35:43–52

Hatagima A, Opromolla DVA, Ura S, Feitosa MF, Beiguelman B, Krieger H (2001). No evidence of linkage between Mitsuda reaction and the *NRAMP1* locus. Int J Lepr Other Mycobact Dis 69:99–103

Hinds D, Risch N (1999). The ASPEX package: affected sib pair exclusion mapping

Holmans P (1993). Asymptotoc properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Holmans P (2002). Detecting gene-gene interactions using affected sib pair analysis with covariates. Hum Hered 53:92–102

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

Holmans P, Craddock N (1997). Efficient strategies for genome scanning using maximum-likelihood affected sib-pair analysis. Am J Hum Genet 60:657–666

Huber P (1967). The behaviour of maximum likelihood estimates under non-standard conditions. In *Proceedings of the fifth Berkley symposium on mathematical statistics and probability*, Volume 1, pp. 221–233. Berkley, CA: University of California press

Izumi S, Sugiyama K, Matsumoto Y, Ohkawa S (1982). Analysis of the immunogenetic background of Japanese leprosy patients by the HLA system. Vox Sang 42:243–7

James J (1971). Frequency in relatives for an all-or-none trait. Ann Hum Genet 35:47–49

Jarvik GP (1998). Complex segregation analyses: uses and limitations. Am J Hum Genet 63:942–946

Job C (1980). Genetics and leprosy. Lepr India 152(3):353–358

Joko S, Numaga J, Kawashima H, Namisato M, Maeda H (2000). Human leukocyte antigens in forms of leprosy among Japanese patients. Int J Lepr Other Mycobact Dis 68(1):49–56

Kaijzel E, Bayley J, van Krugten M, Smith L, van de Linde P, Bakker A (2001). Allele-specific quantification of tumor necrosis factor alpha (TNF) transcription and the role of promoter polymorphisms in rheumatoid arthritis patients and healthy individuals. Genes Immun 2(3):135–44

Karigl G (1982). A mathematical approach to multiple genetic relationships. Theor Popul Biol 21:379–393

Kaur I, Agnihotri N, Mehta M, Dogra S, Ganguly N (2001). Tumor necrosis factor (TNF) production in leprosy patients. Int J Lepr Other Mycobact Dis 69:249–250

Keyu X, de Vries R, Hongming F, van Leeuwen A, Renbiao C, Ganyun Y (1985). HLA-linked control of predisposition to lepromatous leprosy. Int J Lepr Other Mycobact Dis 53(1):56–63

Kim S, Choi I, Dahlberg S, Nisperos B, Kim J, Hansen J (1987). HLA and leprosy in Koreans. Tissue Antigens 29:146–53

Knapp M, Seuchter S, Baur M (1994). Linkage analysis in nulcear families. II. Relationship between affected sib-pair tests and lod score analysis. Hum Hered 44:44–51

Koivisto M, Mannila H (2001). Offspring risk and siblings risk for multilocus traits. Hum Hered 51:209–216

Kong A, Cox N (1997a). Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188

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

Kroeger K, Carville K, Abraham L (1997). The -308 tumor necrosis factor-alpha promoter polymorphism effects transcription. Molecular Immunology 34(5):391–9

Kruglyak L (1997). Nonparametric linkage tests are model free. Am J Hum Genet 61:254–255

Kruglyak L, Daly M, Reeve-Daly M, Lander ES (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Kruglyak L, Lander ES (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lander ES, Kruglyak L (1995). Genetic dissection of complex traits: guidelines for interpretting and reporting linkage results. Nat Genet 11:241–247

Lathrop GM, Lalouel J (1984). Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 36:460–465

Lathrop GM, Lalouel J, Julier C, Ott J (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am J Hum Genet 37:482–498

Lathrop GM, Lalouel J, Ott J (1984). Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446

Levee G, Liu J, Gicquel B, Chanteau S, Schurr E (1994). Genetic control of susceptibility to leprosy in French Polynesia; no evidence for linkage with markers on telomeric human chromosome 2. Int J Lepr Other Mycobact Dis 62(4):499–511

Li H, Huang J (1998). Semiparametric linkage analysis using pseudolikelihoods on neighbouring sets. Ann Hum Genet 62:323–336

Li H, Yang P, Schwartz AG (1998). Analysis of age of onset data from case-control family studies. Biometrics 54:1030–1039

Liang KY, Zeger SL, Qaqish B (1992). Multivariate regression analyses for categorical data. J Roy Stat Soc B 54(1):3–40

Louis E, Franchimont D, Piron A, Gevaert Y, Schaaf-Lafontaine N, Roland S, Mahieu P, Malaise M, Groote DD, Louis R, Belaiche J (1998). Tumor necrosis factor (TNF) gene polymorphism influences TNF-α production in lipopolysaccharide (LPS)-stimulated whole blood cell culture in healthy humans. Clin Exp Immunol 113:401–6

MacLean CJ, Sham PC, Kendler KS (1993). Joint linkage of multiple loci for a complex disorder. Am J Hum Genet 53:353–366

Marquet S, Abel L, Hillaire D, Dessein A (1999). Full results of the genome-wise scan which localises a locus controlling the intensity of infection by *Schistosoma mansoni* on chromosome 5q31–q33. Am J Hum Genet 7:88–97

Marquet S, Abel L, Hillaire D, Dessein H, Kalil J, Feingold J, Weissenbach J, Dessein AJ (1996). Genetic localization of a locus controlling the intensity of infection by *Schistosoma mansoni* on chromosome 5q31–q33. Nat Genet 14:181–184

Marquet S, Schurr E (2001). Genetics of susceptibility to infectious diseases: tuberculosis and leprosy as examples. Drug Metabolism and Disposition 29:479–83

Marsh S, Bodmer J, Albert E, Bodmer W, Bontrop R, Dupont B, Erlich H, Nahsen J, Mach B, Mayr W, Parham P, Petersdorf E, Sasazuki T, Schreuder GT, Strominger J, Sverjgaard A, Terasaki P (2001). Nomenclature for factors of the HLA system, 2000. Tissue Antigens 57:136–283

McCullagh P, Nelder J (1983). Generalized linear models. London: Chapman and Hall

McGuffin P, Huckle P (1990). Simulation of Mendelism revisited: the recessive gene for attending medical school. Am J Hum Genet 46:994–999

Meester SG, MacKay J (1994). A parametric model for cluster correlated categorical data. Biometrics 50(4):954–963

Meisner SJ, Mucklow S, Warner G, Sow SO, Lienhardt C, Hill AV (2001). Association of NRAMP1 polymorphism with leprosy type but not susceptibility to leprosy per se in West Africans. Am J Trop Med Hyg 65:733–735

Meunier F, Philippi A, Martinez M, Demenais F (1997). Affected sib-pair tests for linkage: type I errors within dependent sib-pairs. Genet Epidemiol 14:1107–1111

Mira MT, Alcaïs A, Thuc NV, Thai VH, Huong NT, Ba NN, Verner A, Hudson TJ, Abel L, Schurr E (2003). Chromosome 6q25 is linked to susceptibility to leprosy in a Viewnamese population. Nat Genet 21:412–415

Miyanaga K, Juji T, Maede H, Nakajima S, Kobayashi S (1981). Tuberculoid leprosy and HLA in Japanese. Tissue Antigens 18:331–4

Molenberghs G, Kenward M, Lesaffre E (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. Biometrika 84(1):33–44

Molenberghs G, Lesaffre E (1994). Marginal modeling of correlated ordinal data using a multivariate plackett distribution. J Am Stat Assoc 89(426):633–644

Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill W, Weeks DE (1999). Mega2, a data-handling program for facilitating genetic linkage and association analyses. Am J Hum Genet 65:A436

Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill W, Weeks DE (2001). Mega2 (version 2.3)

Nelsen RB (1999). An introduction to copulas, Volume 139 of *Lecture Notes in Statistics*. New York: Springer-Verlag

Newport M, Blackwell J (1997). Genetic susceptibility to tuberculosis. Balliere's Clinical Infectious Diseases 4(2):207–229

Nicolae DL (1999). Allele sharing models in gene mapping: a likelihood approach. Ph. D. thesis, Department of Statistics, University of Chicago. Chapter 2

O'Connell J, Weeks D (1998). PedCheck: A program for identifying genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266

Olson JM, Cordell HJ (2000). Ascertainment bias in the estimation of sibling genetic risk parameters. Genet Epidemiol 18:217–235

Ott J (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet 51:283–290

Ottenhoff TH, Converse PJ, Bjune G, de Vries RR (1987). HLA antigens and neural reversal reactions in Ethiopian borderline tuberculoid leprosy patients. Int J Lepr Other Mycobact Dis 55(2):261–6

Penrose L (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Annals of Eugenics 8:133–8

Penrose L (1953). The genetical background of common diseases. Acta Genet 4:257–265

Pitchappan R, Kakkanaiah V, Rajashekar R, Arulraj N, Muthukkaruppan V (1984). HLA anitgens in South India. I. Major groups of Tamil Nadu. Tissue Antigens 24:190–196

Pitchappan R, Koteeswaran A, Kakkanaiah V (1989). HLA Bw57 and DR7 association with psoriasis vulgaris in south India. Tissue Antigens 34:133–137

Plackett R (1965). A class of bivariate distributions. J Am Stat Assoc 60:516–522

Pönnighaus J, Fine PE, Bliss L, Sliney I, Bradley D, Rees R (1987). The Lepra Evaluation Project (LEP) and epidemiological study of leprosy in northern Malawi. I: methods. Lepr Rev 52:359–375

Pönnighaus JM, Fine PE, Bliss L (1987). Certainty levels in the diagnosis of leprosy. Int J Lepr Other Mycobact Dis 55(3):454–462

Pönnighaus JM, Fine PE, Sterne JA, Bliss L, Wilson RJ, Malema SS (1994). Incidence rates of leprosy in Karonga District, northern Malawi: patterns by age, sex BCG status and classification. Int J Lepr Other Mycobact Dis 62(1):10–23

Rani R, Fernandez-Vina M, Zaheer S, Beena K, Stastny P (1993). Study of HLA class II alleles by PCR oligotyping in leprosy patients from North India. Tissue Antigens 42:133–137

Rani R, Zaheer S, Mukerjee R (1992). Do human leukocyte antigens have a role to play in differential manifestation of multibacilllary leprosy: a study on multibacillary leprosy patients from North India. Tissue Antigens 40:124–127

Rea TH, Levan NE, Terasaki PI (1976). Histocompatibility antigens in patients with leprosy. J Infect Dis 134(6):615–8

Rice JP, Rochberg N, Neuman RJ, Saccone NL, Liu KY, Zhang X, Culverhouse RC (1999). Covariates in linkage analysis. Genet Epidemiol 17:S691–S695

Risch N (1987). Assessing the role of HLA-linked and unlinked determinants of disease. Am J Hum Genet 40:1–14

Risch N (1990a). Linkage strategies for genetically complex traits. I. Multilocous models. Am J Hum Genet 42:222–228

Risch N (1990b). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 42:229–241

Risch N (1990c). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 42:242–253

Roger M, Levee G, Chanteau S, Gicquel B, Schurr E (1997). No evidence for linkage between leprosy susceptibility and the human natural resistance-associated macrophage protein 1 *NRAMP1* gene in French Polynesia. Int J Lepr Other Mycobact Dis 65(2):197–202

Roy S, Frodsham A, Hazra S, Mascie-Taylor C, Hill AV (1999). Association of vitamin D receptor genotype with leprosy type. J Infect Dis 179:187–191

Roy S, McGuire W, Mascie-Taylor C, Saha B, Hazra S, Hill AV, Kwiatkowski D (1997). Tumor necrosis factor promoter polymorphism and susceptibility to lepromatous leprosy. J Infect Dis 176(2):5

Royall R (1986). Model robust confidence intervals using maximum likelihood estimators. International Statistical Review 54:221–226

Rybicki BA, Elston RC (2000). The relationship between the sibling recurrence-risk ratio and genotype relative risk. Am J Hum Genet 66:593–604

Santos AR, Almeida AS, Suffys PN, Moraes MO, Filho VF, Mattos HJ, Nery JA, Cabello PH, Sampaio EP, Sarno EN (2000). Tumor necrosis factor promoter polymorphism (TNF2) seems to protect against development of severe forms of leprosy in a pilot study in Brazilian patients. Int J Lepr Other Mycobact Dis 68(3):325–7

Santos AR, Suffys PN, Vanderborght PR, Moraes MO, Vieira LM, Cabello PH, Bakker AM, Matos HJ, Huizinga TWJ, Ottenhoff THM, Sampaio EP, Sarno EN (2002). Role of tumor necrosis factor-α and interleukin-10 promoter gene polymorphisms in leprosy. J Infect Dis 186:1687–91

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN, Clayton D (1997). Empirical genomewide significance levels established by whole genome simulations. Genet Epidemiol 14:223–229

Schauf V, Ryan S, Scollard D, Jonasson O, Brown A, Nelson K, Smith T, Vithayasai V (1985). Leprosy associated with HLA-DR2 and DQw1 in the population of northern Thailand. Tissue Antigens 26:243–7

Schliekelman P, Slatkin M (2002). Multiplex relative risk and estimation of the number of loci underlying an inherited disease. Am J Hum Genet 71:1369–1385

Schwartz AG, Yang P, Swanson GM (1996). Familial risk of lung cancer among nonsmokers and their relatives. Am J Epidemiol 144(6):554–562

Sengul H, Weeks DE, Feingold E (2001). A survey of affected-sibship statistics for nonparameteric linkage analysis. Am J Hum Genet 69:179–190

Serjeantson S, Wilson S, Keats B (1979). The genetics of leprosy. Annals of Human Biology 6(4):375–393

Shannon C (1984). A mathematical theory of communication. Bell Syst Tech J 27:379–423

Shaw M, Donaldson I, Collins A, Peacock C, Lins-Lainson Z, Shaw J, Ramos F, Silveira F, Blackwell J (2001). Association and linkage of leprosy phenotypes with HLA class II and tumour necrosis factor genes. Genes Immun 2(4):196–204

Shaw MA, Atkinson S, Dockrell H, Hussain R, Lins-Lainson Z, Shaw J, Ramos F, Silveira F, Mehdi S, Kaukab F, Khaliq S, Chiang T, Blackwell J (1993). An RFLP map for 2q33–q37 from multicase mycobacterial and leishmanial disease families: no evidence for an *Lsh/Ity/Bcg* gene homologue influencing susceptibility to leprosy. Ann Hum Genet 57:251–271

Shields E, Russell D, Pericak-Vance M (1987). Genetic epidemiology of the susceptibility to leprosy. J Clin Invest 79:1139–1143

Shih JH (1998). Modeling multivariate discrete failure time data. Biometrics 54:1115–1128

Shih JH, Louis TA (1995). Inferences on the association parameter in copula models for bivariate survival data. Biometrics 51:1384–1399

Siddiqui MR, Meisner S, Tosh K, Balakrishnan K, Ghei S, Fisher SE, Golding M, Narayan NPS, Sitaraman T, Sengupta U, Pitchappan R, Hill AV (2001). A major susceptibility locus for leprosy in India maps to chromosome 10p13. Nat Genet 27:439–441

Smith G, Walford R, Shephard C, Payne R, Prochazka G (1975). Histocompatability antigens in leprosy. Vox Sang 28:42–49

Somoskovi A, Zissel G, Seitzer U, Gerdes J, Schlaak M, Muller-Quernheim J (1999). Polymorphisms at position -308 in the promoter region of the TNF-alpha and in the first intron of the TNF-beta genes and spontaneous and lipopolysaccharide-induced TNF-alpha release in sarcoidosis. Cytokine 11(11):882–7

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

Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, Mcadams M, Timmerman MM, Brody LC, Tucker MA (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. New Eng J Med 336(20):1401–1408

Stuber F, Udalova I, Book M, Drutskaya L, Kuprash D, Turetskaya R (1995). -308 tumor necrosis factor (TNF) polymorphism is not associated with survival in sever sepsis and is unrelated to lipopolysaccharide inducibility of the human TNF promoter. J Inflamm 46(1):42–50

Suarez B, Eerdewegh PV (1984). A comparison of three affected sib-pair scoring methods to detect HLA-linked disease susceptibility genes. Am J Hum Genet 18:135–146

Suarez B, Hampe C (1994). Linkage and association. Am J Hum Genet 54(3):554–559

Suarez B, Rice J, Reich T (1978). The generalized sib-pair IBD distribution: it's use in the detection of linkage. Ann Hum Genet 42:87–94

Takata H, Sada M, Ozawa S, Sekiguchi S (1978). HLA and mycobaterial infection: increased frequency of B8 in Japanese leprosy. Tissue Antigens 11:61–64

Terwilliger JD, Speer M, Ott J (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 10:217–224

Thompson E (1974). Gene identites and multiple relationships. Biometrics 30:667–680

Tosh K, Meisner S, Siddiqui MR, Balakrishnan K, Ghei S, Golding M, Sengupta U, Pitchappan RM, Hill AV (2002). A region of chromosome 20 is linked to leprosy susceptibility in a South Indian population. J Infect Dis 186:1190–1193

Tosh K, Siddiqui R, Hill A. Association of MRC1 with susceptibility to leprosy in Southern India. In preparation

Trégouët DA, Ducimetière P, Bocquet V, Visvikis S, Soubrier F, Tiret L (1999). A parametric copula model for analysis of familial binary data. Am J Hum Genet 64:886–893

van Eden W, de Vries R, Mehra N, Vaidya M, D'Amaro J, van Rood J (1980). HLA segregation of tuberculoid leprosy: confirmation of the DR2 marker. J Infect Dis 141(6):693–701 van Eden W, Gonzalez NM, de Vries R, Convit J, van Rood J (1985). HLA-linked control of predisposition to lepromatous leprosy. J Infect Dis 151(1):9–14

van Eden W, Mehra N, Vaidya M, Amaro J, Schreuder GT, van Rood J (1981). HLA and sporadic tuberculoid leprosy: a population study in Maharashta. Tissue Antigens 18:189–193

Visentainer J, Tsuneto L, Serra M, Peixoto P, Petzl-Erler M (1997). Association of leprosy with HLA-DR2 in a Southern Brazilian population. Brazilian Journal of Medical and Biological Research 30:51–59

Wacholder S, Hartge P, Strewing JP, Pee D, Mcadams M, Brody L, Tucker M (1998). The kin-cohort study for estimating penetrance. Am J Epidemiol 148(7):623–630

Wang LM, Kimura A, Satoh M, Mineshita S (1999). HLA linked with leprosy in Southern China; HLA-Linked resistance alleles to leprosy. Int J Lepr Other Mycobact Dis 67:403–408

Weeks DE, Lange K (1988). The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–326

Weeks DE, Lange K (1992). A multilocus extension of the affected-pedigree-member method of linkage analysis. Am J Hum Genet 50:859–868

White H (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48:817–830

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

Whittemore AS (1995). Logistic regression of family data from case-control studies. Biometrika 82(1):57–67

Whittemore AS (1996). Genome scanning for linkage: an overview. Am J Hum Genet 59:704–716

Whittemore AS, Halpern J (1994a). A class of tests for linkage using affected pedigree members. Biometrics 50:118–127

Whittemore AS, Halpern J (1994b). Probability of gene identity by descent: computation and applications. Biometrics 50:109–117

Whittemore AS, Tu IP (1998). Simple, robust linkage tests for affected sibs. Am J Hum Genet 62(5):1228–1242

WHO (2002). Leprosy global situation. Weekly Epidemiol Rec 77(1):1–8

Wibawa T, Soebono H, Matsuo M (2002). Association of a missense mutation of the laminin $\alpha 2$ gene with tuberculoid type of leprosy in Indonesian patients. Tropical Medicine and International Health 7(7):631–636

Wilson AG, Symons JA, McDowell TL, McDevitt HO, Duff GW (1997). Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation. Proc Natl Acad Sci USA 94(7):3195–9

Xu J, Myers DA, Pericak-Vance MA (1998). Lod score analysis. In J. L. Haines and M. A. Pericak-Vance (Eds.), *Approaches to gene mapping in complex human diseases*, Chapter 12. New York: Wiley-Liss

Youngchaiyud U, Chandanayingyong D, Vibhatavanija T (1977). The incidence of HLA antigens in leprosy. Vox Sang 32:342–345

APPENDICES

APPENDIX A.

EFFECT OF CENSORING ON θ

In chapter 4, a copula model was proposed for analysis of bivariate present-state disease data, in which a parameter θ measures the degree of association between the disease state of two relatives. It was also acknowledged that censoring of the time individuals remain in a study may bias estimates. It is shown here that if we assume conditional independence, θ is not changed in the presence of censoring. This assumption states that $C_1|T_1$ is conditionally independent of C_2 and T_2 and $C_2|T_1$ is conditionally independent of C_1 and T_1, ie

$$P(C_1, C_2|T_1, T_2) = P(C_1|T_1)P(C_2|T_2).$$

The proof that estimates of θ are not biased in the presence of censoring under this assumption is as follows.

Consider

$$p_{11} = P(T_1 \le t_1, T_2 \le t_2|C_1 > t_1, C_2 > t_2) = \frac{P(C_1 > t_1, C_2 > t_2|T_1 \le t_1, T_2 \le t_2)P(T_1 \le t_1, T_2 \le t_2)}{P(C_1 > t_1, C_2 > t_2)} = \frac{P(C_1 > t_1|T1 \le t_1)P(C_2 > t_2|T_2 \le t_2)P(T_1 \le t_1, T_2 \le t_2)}{P(C_1 > t_1, C_2 > t_2)}$$

Similarly

$$p_{22} = P(T_1 > t_1, T_2 > t_2|C_1 > t_1, C_2 > t_2) = \frac{P(C_1 > t_1|T_1 > t_1)P(C_2 > t_2|T_2 > t_2)P(T_1 > t_1, T_2 > t_2)}{P(C_1 > t_1, C_2 > t_2)},$$

$$p_{12} = P(T_1 \le t_1, T_2 > t_2|C_1 > t_1, C_2 > t_2) = \frac{P(C_1 \le t_1|T_1 > t_1)P(C_2 > t_2|T_2 > t_2)P(T_1 \le t_1, T_2 > t_2)}{P(C_1 > t_1, C_2 > t_2)}, \text{ and}$$

$$p_{21} = P(T_1 > t_1, T_2 > t_2 | C_1 > t_1, C_2 > t_2) = \frac{P(C_1 > t_1 | T_1 > t_1) P(C_2 > t_2 | T_2 \leq t_2) P(T_1 > t_1, T_2 \leq t_2)}{P(C_1 > t_1, C_2 > t_2)}$$

Then

$$\theta = \frac{p_{11} p_{22}}{p_{12} p_{21}} = \frac{P(T_1 \leq t_1, T_2 \leq t_2) P(T_1 > t_1, T_2 > t_2)}{P(T_1 \leq t_1, T_2 > t_2) P(T_1 > t_1, T_2 \leq t_2)}$$

as in equation (4.6).

APPENDIX B.

LIMITS FOR λ_R AS ESTIMATED BY THE COPULA MODEL IN CHAPTER 4

Note that

$$\lambda = \frac{u_1 u_2}{\delta}$$

and

$$\theta = \frac{\delta - u_1\delta - u_2\delta + \delta^2}{u_1 u_2 - u_1\delta - u_2\delta + \delta^2} = \frac{\delta - \epsilon}{u_1 u_2 - \epsilon}$$

where $\epsilon = u\delta + v\delta - \delta^2$. Then if $\lambda > 1$, we have

$$\begin{aligned} \delta &< u_1 u_2 \\ u_1 u_2(\delta - \epsilon) = u_1 u_2 \delta - u_1 u_2 \epsilon &< u_1 u_2 \delta - \delta\epsilon = \delta(u_1 u_2 - \epsilon) \\ \lambda = \frac{u_1 u_2}{\delta} &< \frac{u_1 u_2 - \epsilon}{\delta - \epsilon} = \theta \end{aligned}$$

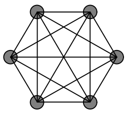
so that $\lambda \in [1, \theta]$. (Conversely, if $\lambda < 1$, then $\lambda \in [\theta, 1]$). λ will approach θ when the disease is rare, ie when u_1, u_2 and δ are small and $\epsilon \to 0$.

Also, when there is no interaction, ie $\delta = u_1 u_2$, then $\theta = 1$ and $\lambda = 1$.

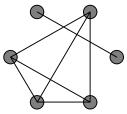
APPENDIX C.

IDENTITY STATES FOR RELATIVE TRIOS

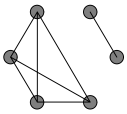
Each horizontal pair of vertices represents the maternal and paternal alleles for each relative. Alleles that are IBD are joined by a solid line.



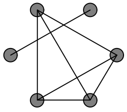
T^*_1



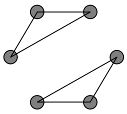
T^*_2



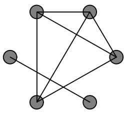
T^*_3



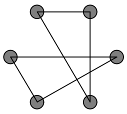
T^*_4



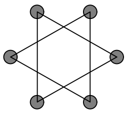
T^*_5



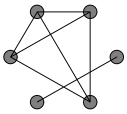
T^*_6



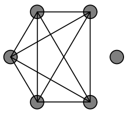
T^*_7



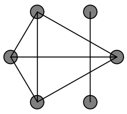
T^*_8



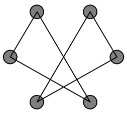
T^*_9



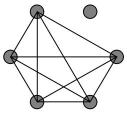
T^*_{10}



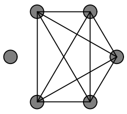
T^*_{11}



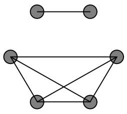
T^*_{12}



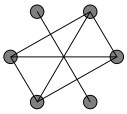
T^*_{13}



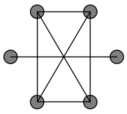
T^*_{14}



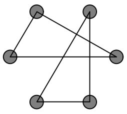
T^*_{15}



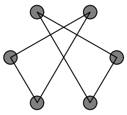
T^*_{16}



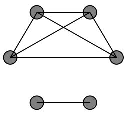
T^*_{17}



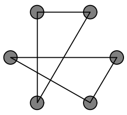
T^*_{18}



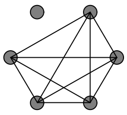
T^*_{19}



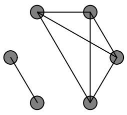
T^*_{20}



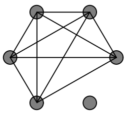
T^*_{21}



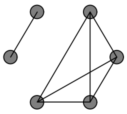
T^*_{22}



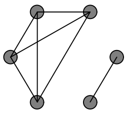
T^*_{23}



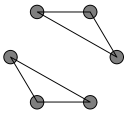
T^*_{24}



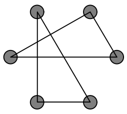
T^*_{25}



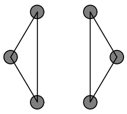
T^*_{26}



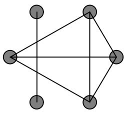
T^*_{27}



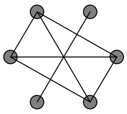
T^*_{28}



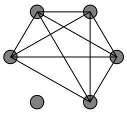
T^*_{29}



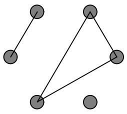
T^*_{30}



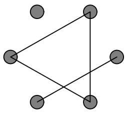
T^*_{31}



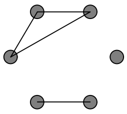
T^*_{32}



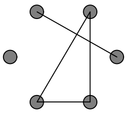
T^*_{33}



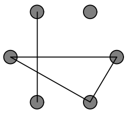
T^*_{34}

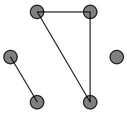


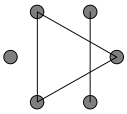
T^*_{35}

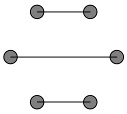


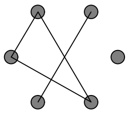
T^*_{36}

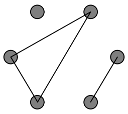

T^*_{37}

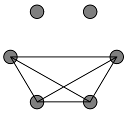

T^*_{38}

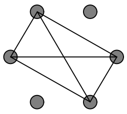

T^*_{39}

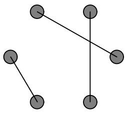

T^*_{40}

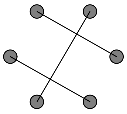

T^*_{41}

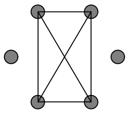

T^*_{42}

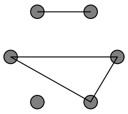

T^*_{43}

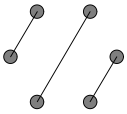

T^*_{44}

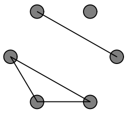

T^*_{45}

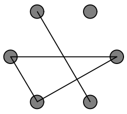

T^*_{46}

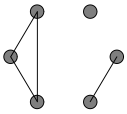

T^*_{47}

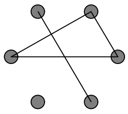

T^*_{48}

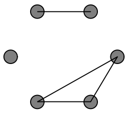

T^*_{49}

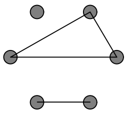

T^*_{50}

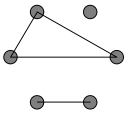

T^*_{51}

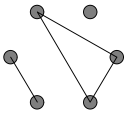

T^*_{52}

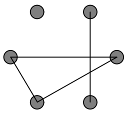

T^*_{53}

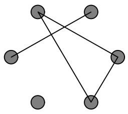

T^*_{54}

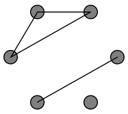

T^*_{55}

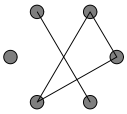

T^*_{56}

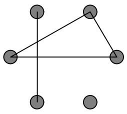

T^*_{57}

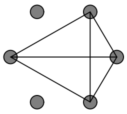

T^*_{58}

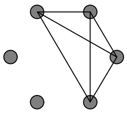

T^*_{59}

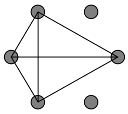

T^*_{60}

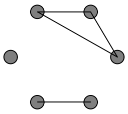

T^*_{61}

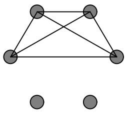

T^*_{62}

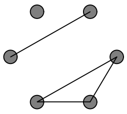

T^*_{63}

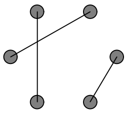

T^*_{64}

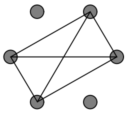

T^*_{65}

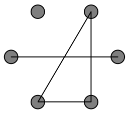

T^*_{66}

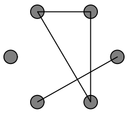

T^*_{67}

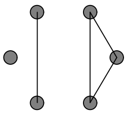

T^*_{68}

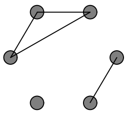

T^*_{69}

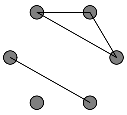

T^*_{70}

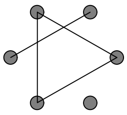

T^*_{71}

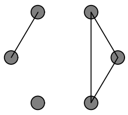

T^*_{72}

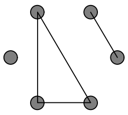

T^*_{73}

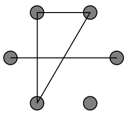

T^*_{74}

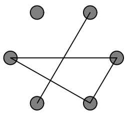

T^*_{75}

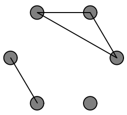

T^*_{76}

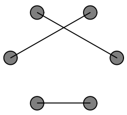

T^*_{77}

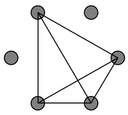

T^*_{78}

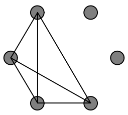

T^*_{79}

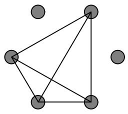

T^*_{80}

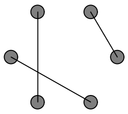

T^*_{81}

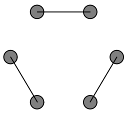

T^*_{82}

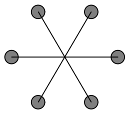

T^*_{83}

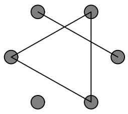

T^*_{84}

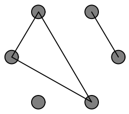

T^*_{85}

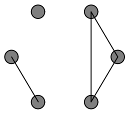

T^*_{86}

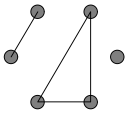

T^*_{87}

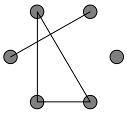

T^*_{88}

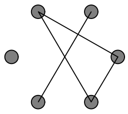

T^*_{89}

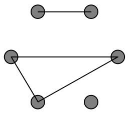

T^*_{90}

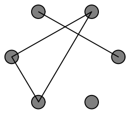

T^*_{91}

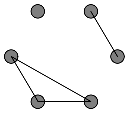

T^*_{92}

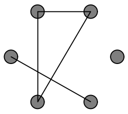

T^*_{93}

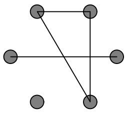

T^*_{94}

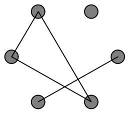

T^*_{95}

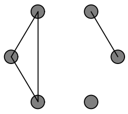

T^*_{96}

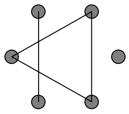

T^*_{97}

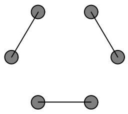

T^*_{98}

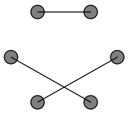

T^*_{99}

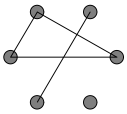

T^*_{100}

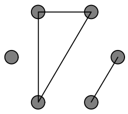

T^*_{101}

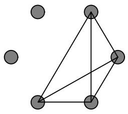

T^*_{102}

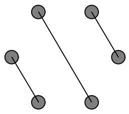

T^*_{103}

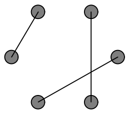

T^*_{104}

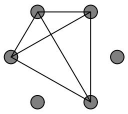

T^*_{105}

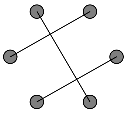

T^*_{106}

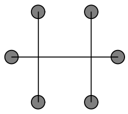

T^*_{107}

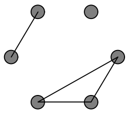

T^*_{108}

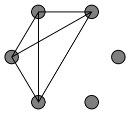

T^*_{109}

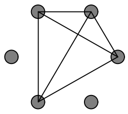

T^*_{110}

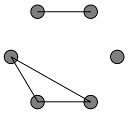

T^*_{111}

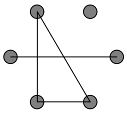

T^*_{112}

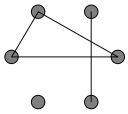

T^*_{113}

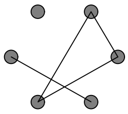

T^*_{114}

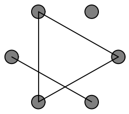

T^*_{115}

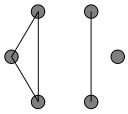

T^*_{116}

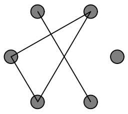

T^*_{117}

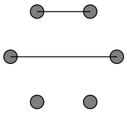

T^*_{118}

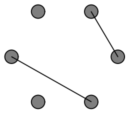

T^*_{119}

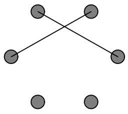

T^*_{120}

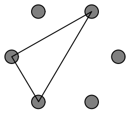

T^*_{121}

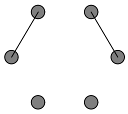

T^*_{122}

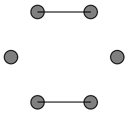

T^*_{123}

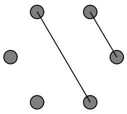

T^*_{124}

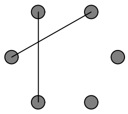

T^*_{125}

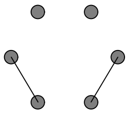

T^*_{126}

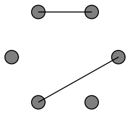

T^*_{127}

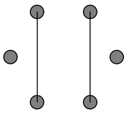

T^*_{128}

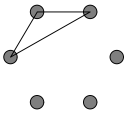

T^*_{129}

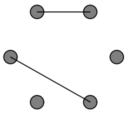

T^*_{130}

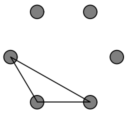

T^*_{131}

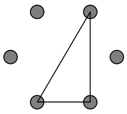

T^*_{132}

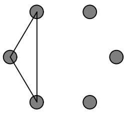

T^*_{133}

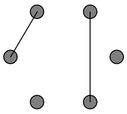

T^*_{134}

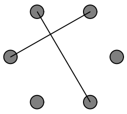

T^*_{135}

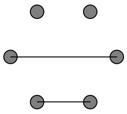

T^*_{136}

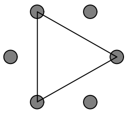

T^*_{137}

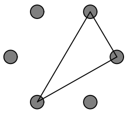

T^*_{138}

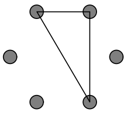

T^*_{139}

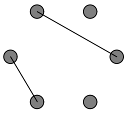

T^*_{140}

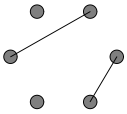

T^*_{141}

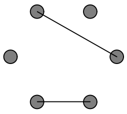

T^*_{142}

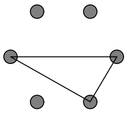

T^*_{143}

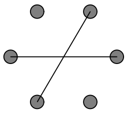

T^*_{144}

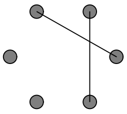

T^*_{145}

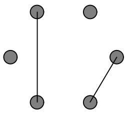

T^*_{146}

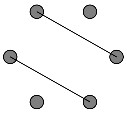

T^*_{147}

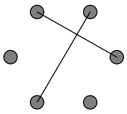

T^*_{148}

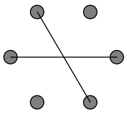

T^*_{149}

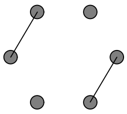

T^*_{150}

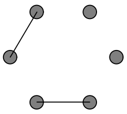

T^*_{151}

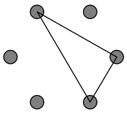

T^*_{152}

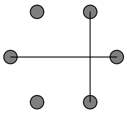

T^*_{153}

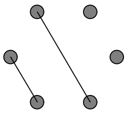

T^*_{154}

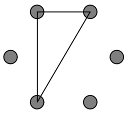

T^*_{155}

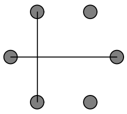

T^*_{156}

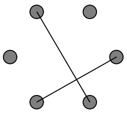

T^*_{157}

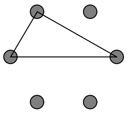

T^*_{158}

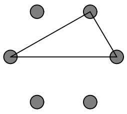

T^*_{159}

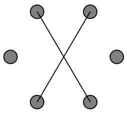

T^*_{160}

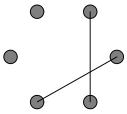

T^*_{161}

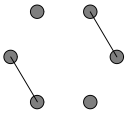

T^*_{162}

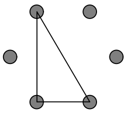

T^*_{163}

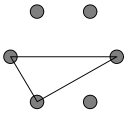

T^*_{164}

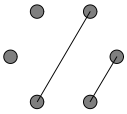

T^*_{165}

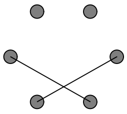

T^*_{166}

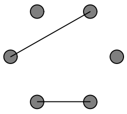

T^*_{167}

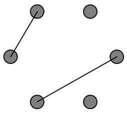

T^*_{168}

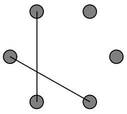

T^*_{169}

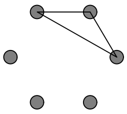

T^*_{170}

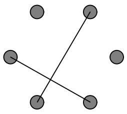

T^*_{171}

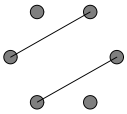

T^*_{172}

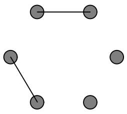

T^*_{173}

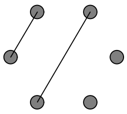

T^*_{174}

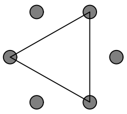

T^*_{175}

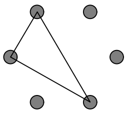

T^*_{176}

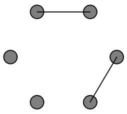

T^*_{177}

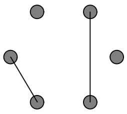

T^*_{178}

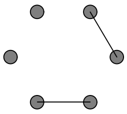

T^*_{179}

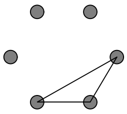

T^*_{180}

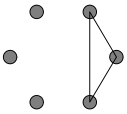

T^*_{181}

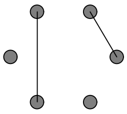

T^*_{182}

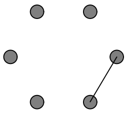

T^*_{183}

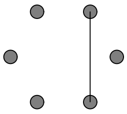

T^*_{184}

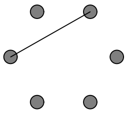

T^*_{185}

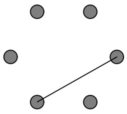

T^*_{186}

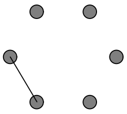

T^*_{187}

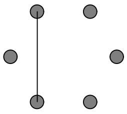

T^*_{188}

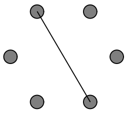

T^*_{189}

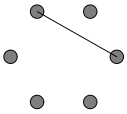

T^*_{190}

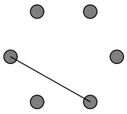

T^*_{191}

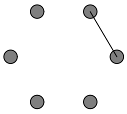

T^*_{192}

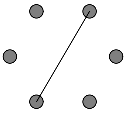

T^*_{193}

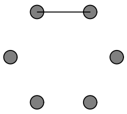

T^*_{194}

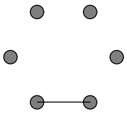

T^*_{195}

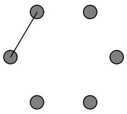

T^*_{196}

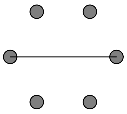

T^*_{197}

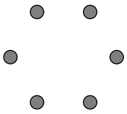

T^*_{198}


T^*_{199}


T^*_{200}


T^*_{201}


T^*_{202}


T^*_{203}

APPENDIX D.

SUMMARY STATISTICS FOR MARKERS USED IN LINKAGE ANALYSIS

Let a given locus have n alleles, and let the population frequency of allele i be π_i. Then a common measure of marker polymorphism is

Definition D.1: the *heterozygosity* of locus, defined as the probability that an individual selected at random from the population is heterozygous:

$$1 - \sum_{i=1}^{n} \pi_i^2.$$

The markers used in stage I of the genome screen described in chapter 6 are listed below, with their position, the number of alleles found and the single-point MLS scores.

Marker	Position (cM)	# alleles	Heterozygosity	Singlepoint MLS score
D10S249	0.0	6	0.71	0.00
D10S591	12.3	8	0.76	0.37
D10S189	17.3	5	0.63	0.55
D10S547	28.1	13	0.85	0.71
D10S1653	38.8	9	0.63	0.95
D10S548	43.4	6	0.52	0.24
D10S197	50.5	12	0.85	0.52
D10S208	60.2	9	0.86	0.09
D10S196	72.5	14	0.82	0.95
D10S1652	83.3	13	0.85	0.00
D10S537	93.8	13	0.85	0.22
D10S1686	109.2	19	0.90	0.05
D10S185	123.3	15	0.85	0.35
D10S192	131.2	13	0.66	0.09
D10S597	137.6	9	0.76	0.00
D10S1693	146.1	20	0.86	0.20
D10S587	156.6	14	0.84	0.49
D10S1651	178.3	14	0.87	0.46

Table D.1: Summary statistics for markers used in stage I of genome screen and single-point MLS scores

Marker	Position (cM)	# alleles	Heterozygosity	Singlepoint MLS score
D10S212	180.7	7	0.64	0.07
D15S128	6.1	10	0.79	0.17
D15S1002	14.5	12	0.83	0.00
D15S165	20.2	15	0.89	0.00
D15S1007	25.9	18	0.88	0.16
D15S1012	35.3	12	0.82	0.00
D15S994	40.0	14	0.89	0.47
D15S978	45.5	13	0.80	0.01
D15S117	50.8	14	0.87	0.00
D15S153	62.1	12	0.81	0.00
D15S131	70.7	16	0.86	0.00
D15S205	77.4	18	0.90	0.00
D15S130	98.0	9	0.76	0.00
D15S120	109.6	11	0.87	0.08
D20S115	20.9	10	0.72	0.01
D20S832	84.0	7	0.76	0.02
D20S102	85.8	5	0.46	0.35
D20S171	94.4	15	0.87	0.14
D20S173	96.5	13	0.79	0.08
D21S1256	8.6	14	0.85	0.01
D21S1914	23.0	13	0.83	0.03
D21S263	31.4	15	0.87	0.19
D21S1252	38.7	15	0.89	0.03
D21S266	49.9	11	0.73	**1.39**
DXS1060	10.1	11	0.75	0.01
DXS8051	15.7	14	0.89	0.00
DXS987	25.5	13	0.85	0.03
DXS1226	36.8	15	0.88	0.39
DXS1214	46.2	9	0.81	0.35
DXS1068	56.2	11	0.83	0.01
DXS993	66.1	11	0.77	**1.06**
DXS991	86.9	12	0.82	0.50
DXS986	95.9	15	0.88	0.06
DXS990	104.9	10	0.76	0.12
DXS1106	115.1	8	0.79	0.30
DXS8055	126.8	6	0.75	0.76
DXS1001	139.4	9	0.81	0.34
DXS1047	150.3	15	0.87	0.06
DXS1227	164.7	13	0.81	0.06
DXS8043	176.7	12	0.81	0.03

Table D.1: Summary statistics for markers used in stage I of genome screen and single-point MLS scores

Marker	Position (cM)	# alleles	Heterozygosity	Singlepoint MLS score
DXS8091	186.3	15	0.85	0.06
DXS1073	196.5	14	0.77	**2.09**
D5S1981	0.6	11	0.79	0.02
D5S406	10.7	11	0.83	0.31
D5S630	18.6	23	0.92	0.53
D5S416	27.9	12	0.83	0.02
D5S419	39.5	11	0.85	0.57
D5S426	51.6	10	0.86	0.35
D5S418	58.1	7	0.79	0.00
D5S647	74.7	15	0.82	0.09
D5S424	82.8	10	0.73	0.26
D5S641	92.3	15	0.87	0.19
D5S428	95.4	8	0.79	0.59
D5S644	104.5	15	0.86	**1.17**
D5S433	112.2	12	0.83	0.60
D5S2027	118.9	8	0.79	0.03
D5S471	129.6	10	0.80	0.11
D5S2115	138.6	16	0.87	0.07
D5S410	156.0	7	0.51	0.12
D5S422	163.9	14	0.89	0.01
D5S400	174.3	16	0.90	0.22
D5S408	195.8	13	0.84	0.00
D6S1574	8.7	15	0.81	0.00
D6S309	13.6	11	0.85	0.00
D6S470	17.7	10	0.76	0.00
D6S289	29.6	11	0.79	0.00
D6S422	35.7	14	0.89	0.00
D6S1610	53.9	13	0.80	0.00
D6S257	80.0	19	0.91	0.00
D6S460	90.0	13	0.88	0.00
D6S462	99.0	13	0.80	0.31
D6S434	109.2	15	0.72	0.43
D6S287	122.0	8	0.80	0.00
D6S262	129.8	10	0.87	0.11
D6S292	138.2	12	0.88	0.80
D6S308	145.5	7	0.52	0.00
D6S441	155.3	16	0.88	0.08
D6S1581	165.0	13	0.77	0.22
D6S264	179.1	12	0.85	0.24
D6S446	188.4	6	0.69	0.63

Table D.1: Summary statistics for markers used in stage I of genome screen and single-point MLS scores

Marker	Position (cM)	# alleles	Heterozygosity	Singlepoint MLS score
D6S281	201.1	11	0.79	0.51
D9S288	8.8	16	0.87	0.00
D9S286	16.8	16	0.87	0.00
D9S285	27.9	14	0.86	0.11
D9S157	31.8	14	0.80	0.00
D9S171	42.0	11	0.64	0.00
D9S1817	57.9	15	0.82	0.00
D9S273	64.5	13	0.83	0.00
D9S175	68.8	19	0.85	0.00
D9S167	82.4	12	0.78	0.00
D9S283	93.2	10	0.76	0.00
D9S1690	106.5	9	0.71	0.00
D9S1677	117.8	13	0.84	0.00
D9S1776	124.2	14	0.83	0.00
D9S1682	132.9	8	0.72	0.53
D9S290	141.1	17	0.82	0.00
D9S164	148.1	13	0.82	0.38
D9S1826	160.2	10	0.70	0.60
D9S158	163.0	13	0.87	0.71

Table D.1: Summary statistics for markers used in stage I of genome screen and single-point MLS scores

APPENDIX E.
CHROMOSOME 21 MARKER MAP ORDER

E.1 Published genethon map

Position (cM)	Marker
23.0	D21S1914
31.4	D21S263
38.7	D21S1252
41.8	D21S267
42.4	D21S1891
42.8	D21S1255
48.1	D21S1893
49.9	D21S266
50.7	D21S1906
51.6	D21S1260
57.7	D21S1890
57.8	D21S1885
58.3	D21S1912
58.9	D21S1903
59.6	D21S1897
63.5	D21S2057

E.2 Results of SIMWALK2

```
 POSITION   MARKER     RECOMB.   RECOMBINATION EVENTS   SIGNIFICANCE
Haldane cM   NAME     FRACTION    OBSERVED & EXPECTED     (P-VALUE)

    0.000   D21S1914
                       0.07732     125.292    126.032      0.54063
    8.400   D21S263
                       0.06792     130.743    110.710      0.02953
   15.700   D21S1252
                       0.03006      41.634     48.998      0.87472
   18.800   D21S267
                       0.00596      28.889      9.715      0.00000  !##
   19.399   D21S1891
                       0.00398       5.822      6.487      0.65712
   19.799   D21S1255
                       0.05029      86.734     81.973      0.30997
   25.099   D21S1893
                       0.01768      39.751     28.818      0.02951
   26.899   D21S266
                       0.00794       9.358     12.942      0.87688
   27.699   D21S1906
                       0.00892      15.685     14.540      0.41608
   28.600   D21S1260
                       0.05743     116.290     93.611      0.01082
   34.700   D21S1890
                       0.00100       2.290      1.630      0.39773
   34.800   D21S1885
                       0.00498      13.904      8.117      0.03978
   35.301   D21S1912
                       0.00596      11.367      9.715      0.33857
   35.900   D21S1903
                       0.00695      13.482     11.329      0.29804
   36.600   D21S1897
                       0.03752      62.528     61.158      0.44699
   40.500   D21S2057


!## indicates a p-value which is so small that one should reconsider whether
    the specified recombination fraction for this interval is too small!
```

E.3 Results of sib_map

The results of sib_map run using nuclear families and the stage II chromosome 21 data to check map order are below. To aid interpretation, they are prefaced with a quote from the ASPEX (Hinds and Risch, 1999) manual.

> The normal output of sib_map consists of the two-point and multipoint distances between each pair of adjacent markers, and the corresponding support intervals. If two markers are determined to be unlinked, their distance will be reported as "[inf]" (for "infinity"). In this case, the support interval will give a lower bound on the most likely distance between the pair. A LOD score is reported for each interval, giving the likelihood for the most likely distance, compared to the likelihood of the two markers being unlinked.
>
> From simulations, we estimate that for support levels of 0.2, 0.6, and 0.8 LOD units, the true distance should be within the support interval about 70%, 90%, and 95% of the time, respectively. These are not strict confidence intervals, however, so these probabilities should be used only as rough guidelines.
>
> If verbose (-v) output is selected, then tables of LOD scores versus distance for each marker pair will also be generated.
>
> If do_shuffle is true, then the output is, for each marker, a table giving three-point distance estimates for that marker with every other pair of adjacent markers along the map. The total distance spanning the three markers, and the corresponding LOD score, is generated for all possible orders of markers (XAB, AXB, ABX). Thus, a comparison of the LOD scores indicates where the test marker is likely to be in relation to the pair.
>
> Following the distances and LOD scores, sib_map will print one of several symbols based on a comparison of the LOD scores. If the test marker appears to be to the left of the specified pair, then "<" or "<<" will be printed: the number of arrows indicates that the LOD score difference exceeds that number times the value of support. Similarly, ">" or ">>" will be printed if the marker is to the right of the pair. "+" or "++" will be printed if the marker is most likely to be between the specified pair. If the map order is correct and well supported by the data, the symbols for each marker should show a pattern of ">" rows, then one "+" row, then "<" rows, as the marker is shuffled through its true position. To use the output to position new loci on a map, in the parameter file for sib_map, list the new loci first in the map, followed by all the already-mapped loci in their proper order.

```
Three-point distances for D21S1914:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
 D21S263 -- D21S1252  0.184  76.84  0.215  65.22  0.258  56.74  <<
D21S1252 -- D21S267   0.181  81.67  0.243  41.61  0.175  82.50  >
 D21S267 -- D21S1891  0.247  56.90  0.295  28.63  0.251  56.76
D21S1891 -- D21S1255  0.223  98.01  0.261  34.60  0.223  98.03
D21S1255 -- D21S1893  0.290  54.96  0.410  20.86  0.348  49.84  <<
D21S1893 -- D21S266   0.379  41.17  0.629   8.47  0.410  40.00  <
 D21S266 -- D21S1906  0.402  25.86  0.630   7.00  0.343  27.67  >>
D21S1906 -- D21S1260  0.315  34.45  0.513  10.54  0.357  33.18  <<
D21S1260 -- D21S1890  0.439  44.20  0.770   6.21  0.733  40.24  <<
D21S1890 -- D21S1885  0.628  76.64  0.925   2.20  0.613  76.72
D21S1885 -- D21S1912  0.592  47.64  0.778   2.91  0.654  47.28
D21S1912 -- D21S1903  0.628  62.35  0.965   1.39  0.645  62.29
D21S1903 -- D21S1897  0.652  56.49  1.267   0.69  0.660  56.44
D21S1897 -- D21S2057  0.748   6.02  1.103   0.51  0.726   5.96

Three-point distances for D21S263:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S1252  0.215  65.22  0.184  76.84  0.258  56.74  ++
D21S1252 -- D21S267   0.108 103.08  0.164  73.88  0.107 103.66
 D21S267 -- D21S1891  0.170  74.26  0.223  52.66  0.178  70.60  <<
D21S1891 -- D21S1255  0.156 111.29  0.226  53.95  0.151 112.35  >
D21S1255 -- D21S1893  0.215  69.90  0.280  45.35  0.261  60.13  <<
D21S1893 -- D21S266   0.275  50.29  0.455  20.46  0.280  48.21  <<
 D21S266 -- D21S1906  0.297  31.65  0.424  17.90  0.263  34.88  >>
D21S1906 -- D21S1260  0.204  45.42  0.319  25.03  0.219  43.53  <<
D21S1260 -- D21S1890  0.319  51.56  0.474  18.15  0.449  44.53  <<
D21S1890 -- D21S1885  0.369  81.26  0.555  11.38  0.372  81.11
D21S1885 -- D21S1912  0.372  51.54  0.582   8.71  0.381  51.13
D21S1912 -- D21S1903  0.380  66.18  0.590   8.67  0.390  65.96
D21S1903 -- D21S1897  0.386  60.61  0.601   9.41  0.388  60.39
D21S1897 -- D21S2057  0.404   9.56  0.632   4.34  0.378   8.94  <

Three-point distances for D21S1252:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.215  65.22  0.258  56.74  0.184  76.84  >>
 D21S263 -- D21S267   0.164  73.88  0.108 103.08  0.107 103.66
 D21S267 -- D21S1891  0.057 125.53  0.056 124.92  0.077 111.05  <
D21S1891 -- D21S1255  0.052 151.94  0.065 130.87  0.050 154.05  >>
D21S1255 -- D21S1893  0.114 110.46  0.136 100.01  0.148  88.44  <<
D21S1893 -- D21S266   0.158  75.57  0.213  59.25  0.152  72.46  <<
 D21S266 -- D21S1906  0.168  50.98  0.228  43.14  0.164  51.93  >
D21S1906 -- D21S1260  0.135  61.26  0.191  50.67  0.141  61.31
D21S1260 -- D21S1890  0.216  69.53  0.285  46.96  0.317  54.23  <<
D21S1890 -- D21S1885  0.240  91.89  0.356  31.01  0.235  92.61  >
```

```
D21S1885 -- D21S1912  0.241  62.02  0.356  28.16  0.247  60.85  <
D21S1912 -- D21S1903  0.296  71.87  0.429  23.66  0.300  71.73
D21S1903 -- D21S1897  0.310  66.02  0.513  16.38  0.305  66.09
D21S1897 -- D21S2057  0.420   9.94  0.885   4.27  0.421   8.36  <<

Three-point distances for D21S267:

                          [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.210  56.81  0.195  51.55  0.177  66.47  >>
 D21S263 -- D21S1252  0.164  73.88  0.107 103.66  0.108 103.08
D21S1252 -- D21S1891  0.056 124.92  0.057 125.53  0.077 111.05  +
D21S1891 -- D21S1255  0.044 142.54  0.043 130.31  0.041 144.38  >>
D21S1255 -- D21S1893  0.100 101.07  0.088  98.23  0.120  83.65  <<
D21S1893 -- D21S266   0.144  67.85  0.168  54.46  0.158  62.56  <<
 D21S266 -- D21S1906  0.155  45.03  0.168  43.97  0.140  49.63  >>
D21S1906 -- D21S1260  0.101  62.08  0.116  56.63  0.116  57.82  <<
D21S1260 -- D21S1890  0.215  62.91  0.240  42.97  0.346  49.72  <<
D21S1890 -- D21S1885  0.288  84.46  0.323  23.99  0.280  85.00
D21S1885 -- D21S1912  0.261  56.47  0.317  20.75  0.270  55.51  <
D21S1912 -- D21S1903  0.308  68.84  0.379  19.16  0.310  68.88
D21S1903 -- D21S1897  0.339  61.87  0.534  10.45  0.347  61.37
D21S1897 -- D21S2057  0.510   7.30  0.671   2.16  0.495   6.95

Three-point distances for D21S1891:

                          [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.282  50.16  0.323  32.56  0.255  55.32  >>
 D21S263 -- D21S1252  0.227  60.10  0.170  81.47  0.159  88.74  >>
D21S1252 -- D21S267   0.056 124.92  0.077 111.05  0.057 125.53  >
 D21S267 -- D21S1255  0.043 130.31  0.044 142.54  0.041 144.38  >>
D21S1255 -- D21S1893  0.069 139.77  0.070 139.97  0.111 105.40
D21S1893 -- D21S266   0.128  85.29  0.143  78.22  0.107  84.00  <<
 D21S266 -- D21S1906  0.132  54.00  0.168  47.02  0.129  53.63
D21S1906 -- D21S1260  0.102  68.58  0.133  64.43  0.114  70.81  >>
D21S1260 -- D21S1890  0.172  82.53  0.217  66.14  0.258  60.85  <<
D21S1890 -- D21S1885  0.201  94.28  0.283  41.22  0.196  95.11  >
D21S1885 -- D21S1912  0.194  65.27  0.269  35.61  0.206  62.55  <<
D21S1912 -- D21S1903  0.251  73.75  0.354  28.29  0.251  73.93
D21S1903 -- D21S1897  0.269  67.28  0.453  19.45  0.279  66.48  <
D21S1897 -- D21S2057  0.358  11.76  0.664   6.42  0.309  11.30

Three-point distances for D21S1255:

                          [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.249  55.91  0.280  44.86  0.212  65.43  >>
 D21S263 -- D21S1252  0.184  73.94  0.148  99.70  0.139 103.54  >>
D21S1252 -- D21S267   0.043 136.99  0.053 131.63  0.042 138.69  >>
 D21S267 -- D21S1891  0.043 130.31  0.041 144.38  0.044 142.54  ++
D21S1891 -- D21S1893  0.070 139.97  0.069 139.77  0.111 105.40
```

```
D21S1893 -- D21S266   0.140  82.69  0.160  73.71  0.119  82.01  <
 D21S266 -- D21S1906  0.139  55.33  0.170  48.17  0.132  55.74
D21S1906 -- D21S1260  0.117  65.49  0.146  59.78  0.130  66.80  >>
D21S1260 -- D21S1890  0.188  77.54  0.241  59.33  0.272  59.27  <<
D21S1890 -- D21S1885  0.207  94.19  0.300  36.88  0.203  94.93  >
D21S1885 -- D21S1912  0.202  65.35  0.278  35.74  0.196  65.19
D21S1912 -- D21S1903  0.220  77.54  0.314  34.47  0.227  76.78  <
D21S1903 -- D21S1897  0.244  70.36  0.376  24.59  0.248  69.98
D21S1897 -- D21S2057  0.328  12.84  0.713   7.17  0.298  11.35  <<

Three-point distances for D21S1893:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.349  44.54  0.414  21.93  0.290  50.36  >>
 D21S263 -- D21S1252  0.270  53.43  0.258  54.28  0.227  67.03  >>
D21S1252 -- D21S267   0.111  98.06  0.140  74.41  0.109  98.81  >
 D21S267 -- D21S1891  0.097  87.05  0.115  86.25  0.110  93.57  >>
D21S1891 -- D21S1255  0.070 139.97  0.111 105.40  0.069 139.77
D21S1255 -- D21S266   0.160  73.71  0.140  82.69  0.119  82.01  +
 D21S266 -- D21S1906  0.084  70.42  0.088  69.12  0.076  69.43  <
D21S1906 -- D21S1260  0.076  76.79  0.088  75.58  0.082  78.87  >>
D21S1260 -- D21S1890  0.128  93.48  0.152  83.13  0.195  73.31  <<
D21S1890 -- D21S1885  0.135 107.20  0.172  66.25  0.137 106.72
D21S1885 -- D21S1912  0.138  76.29  0.207  47.58  0.147  73.53  <<
D21S1912 -- D21S1903  0.174  84.29  0.239  47.22  0.172  84.81
D21S1903 -- D21S1897  0.185  78.08  0.270  41.81  0.186  77.62
D21S1897 -- D21S2057  0.249  19.37  0.301  17.60  0.201  18.67  <

Three-point distances for D21S266:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.487  40.37  0.569  10.17  0.399  43.24  >>
 D21S263 -- D21S1252  0.344  45.04  0.305  35.62  0.267  56.05  >>
D21S1252 -- D21S267   0.150  86.22  0.187  51.04  0.151  85.88
 D21S267 -- D21S1891  0.159  68.41  0.181  51.46  0.147  72.00  >>
D21S1891 -- D21S1255  0.104 117.22  0.124  74.45  0.103 117.31
D21S1255 -- D21S1893  0.160  73.71  0.119  82.01  0.140  82.69  >
D21S1893 -- D21S1906  0.088  69.12  0.084  70.42  0.076  69.43  +
D21S1906 -- D21S1260  0.047  85.38  0.046 101.32  0.046 101.32
D21S1260 -- D21S1890  0.087 117.64  0.087 117.64  0.133  87.65
D21S1890 -- D21S1885  0.076 122.24  0.081  96.27  0.079 121.16  <
D21S1885 -- D21S1912  0.109  82.94  0.130  67.61  0.115  80.69  <<
D21S1912 -- D21S1903  0.104  97.44  0.123  73.97  0.110  96.37  <
D21S1903 -- D21S1897  0.107  93.10  0.120  71.02  0.109  91.04  <<
D21S1897 -- D21S2057  0.186  24.10  0.220  22.44  0.141  23.68

Three-point distances for D21S1906:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
```

```
D21S1914 -- D21S263   0.307  44.38  0.254  25.96  0.264  47.78  >>
 D21S263 -- D21S1252  0.262  48.16  0.185  48.11  0.200  57.21  >>
D21S1252 -- D21S267   0.096  85.97  0.089  69.40  0.093  86.70  >
 D21S267 -- D21S1891  0.102  69.52  0.099  62.06  0.130  64.35  <<
D21S1891 -- D21S1255  0.091 108.57  0.075  77.40  0.087 109.27  >
D21S1255 -- D21S1893  0.140  65.32  0.106  66.21  0.114  71.18  >>
D21S1893 -- D21S266   0.088  69.12  0.076  69.43  0.084  70.42  >
 D21S266 -- D21S1260  0.046 101.32  0.047  85.38  0.046 101.32
D21S1260 -- D21S1890  0.139  69.58  0.134  62.80  0.235  54.40  <<
D21S1890 -- D21S1885  0.162  90.89  0.131  49.62  0.161  90.86
D21S1885 -- D21S1912  0.162  60.94  0.148  37.57  0.190  58.34  <<
D21S1912 -- D21S1903  0.211  72.44  0.182  35.67  0.213  72.44
D21S1903 -- D21S1897  0.238  65.27  0.231  25.75  0.241  64.75
D21S1897 -- D21S2057  0.448   7.52  0.422   3.59  0.382   7.70

Three-point distances for D21S1260:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.371  44.05  0.455  18.77  0.316  48.12  >>
 D21S263 -- D21S1252  0.289  50.81  0.294  44.82  0.242  61.98  >>
D21S1252 -- D21S267   0.133  92.27  0.159  65.28  0.131  92.56
 D21S267 -- D21S1891  0.113  84.07  0.137  74.33  0.114  89.67  >>
D21S1891 -- D21S1255  0.087 131.76  0.109  94.06  0.087 131.56
D21S1255 -- D21S1893  0.138  86.97  0.126  89.69  0.130  92.91  >>
D21S1893 -- D21S266   0.084  96.73  0.055 122.14  0.055 122.14
 D21S266 -- D21S1906  0.046 101.32  0.046 101.32  0.047  85.38
D21S1906 -- D21S1890  0.134  62.80  0.139  69.58  0.235  54.40  ++
D21S1890 -- D21S1885  0.087 121.68  0.101  91.88  0.086 121.67
D21S1885 -- D21S1912  0.094  86.05  0.122  68.45  0.109  80.33  <<
D21S1912 -- D21S1903  0.114  97.07  0.141  71.13  0.115  97.56
D21S1903 -- D21S1897  0.123  91.02  0.184  58.58  0.129  88.10  <<
D21S1897 -- D21S2057  0.206  24.04  0.223  22.25  0.169  22.51  <<

Three-point distances for D21S1890:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.636  38.27  0.996   4.17  0.468  41.10  >>
 D21S263 -- D21S1252  0.415  41.90  0.523  18.76  0.370  46.72  >>
D21S1252 -- D21S267   0.246  76.26  0.460  22.68  0.247  75.91
 D21S267 -- D21S1891  0.248  60.17  0.413  28.96  0.230  64.60  >>
D21S1891 -- D21S1255  0.184 109.51  0.331  39.53  0.183 109.44
D21S1255 -- D21S1893  0.236  65.46  0.293  49.61  0.208  73.95  >>
D21S1893 -- D21S266   0.146  78.46  0.176  73.52  0.133  83.90  >>
 D21S266 -- D21S1906  0.153  62.40  0.213  54.30  0.142  54.70  <<
D21S1906 -- D21S1260  0.134  62.80  0.235  54.40  0.139  69.58  >>
D21S1260 -- D21S1885  0.101  91.88  0.087 121.68  0.086 121.67
D21S1885 -- D21S1912  0.023 142.66  0.022 143.18  0.031 133.10
D21S1912 -- D21S1903  0.025 155.27  0.035 149.38  0.027 155.58
D21S1903 -- D21S1897  0.040 146.86  0.048 142.65  0.041 141.79  <<
```

```
D21S1897 -- D21S2057  0.116  62.59  0.109  63.40  0.054  56.30  +

Three-point distances for D21S1885:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.504  39.57  0.618   8.28  0.398  42.45  >>
 D21S263 -- D21S1252  0.346  43.85  0.386  24.48  0.314  48.90  >>
D21S1252 -- D21S267   0.205  76.91  0.276  31.93  0.209  76.10  <
 D21S267 -- D21S1891  0.200  60.19  0.268  32.86  0.184  64.18  >>
D21S1891 -- D21S1255  0.144 108.18  0.187  49.29  0.145 107.94
D21S1255 -- D21S1893  0.219  61.86  0.230  46.10  0.204  65.53  >>
D21S1893 -- D21S266   0.143  69.64  0.149  64.82  0.123  76.62  >>
 D21S266 -- D21S1906  0.134  57.48  0.156  51.06  0.135  50.05  <<
D21S1906 -- D21S1260  0.122  55.32  0.145  51.51  0.116  61.23  >>
D21S1260 -- D21S1890  0.101  91.88  0.086 121.67  0.087 121.68
D21S1890 -- D21S1912  0.022 143.18  0.023 142.66  0.031 133.10
D21S1912 -- D21S1903  0.029 131.16  0.039 123.84  0.028 131.13
D21S1903 -- D21S1897  0.042 122.58  0.039 120.43  0.043 119.60  <<
D21S1897 -- D21S2057  0.105  42.74  0.091  43.67  0.057  38.69  +

Three-point distances for D21S1912:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.564  38.22  0.609   5.07  0.428  40.19  >>
 D21S263 -- D21S1252  0.393  40.04  0.403  20.09  0.314  47.07  >>
D21S1252 -- D21S267   0.206  74.89  0.282  29.65  0.209  74.26  <
 D21S267 -- D21S1891  0.212  57.61  0.305  26.42  0.202  59.94  >>
D21S1891 -- D21S1255  0.155 104.34  0.196  48.71  0.155 104.41
D21S1255 -- D21S1893  0.213  61.86  0.247  41.95  0.212  62.05
D21S1893 -- D21S266   0.171  61.39  0.195  49.45  0.155  65.19  >>
 D21S266 -- D21S1906  0.162  47.08  0.208  37.94  0.153  43.97  <<
D21S1906 -- D21S1260  0.152  48.51  0.212  37.80  0.152  52.29  >>
D21S1260 -- D21S1890  0.138  77.54  0.118  93.71  0.109  99.42  >>
D21S1890 -- D21S1885  0.022 143.18  0.031 133.10  0.023 142.66
D21S1885 -- D21S1903  0.039 123.84  0.029 131.16  0.028 131.13
D21S1903 -- D21S1897  0.023 127.85  0.024 126.31  0.023 125.66  <<
D21S1897 -- D21S2057  0.112  42.00  0.106  41.97  0.050  36.74

Three-point distances for D21S1903:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.588  38.76  0.887   5.38  0.447  41.62  >>
 D21S263 -- D21S1252  0.410  42.10  0.510  17.56  0.390  44.89  >>
D21S1252 -- D21S267   0.264  73.52  0.482  19.45  0.266  73.25
 D21S267 -- D21S1891  0.276  56.52  0.459  20.37  0.255  59.38  >>
D21S1891 -- D21S1255  0.214 102.70  0.372  29.34  0.216 102.52
D21S1255 -- D21S1893  0.272  59.62  0.320  39.32  0.238  65.33  >>
D21S1893 -- D21S266   0.192  65.41  0.216  60.47  0.168  72.78  >>
 D21S266 -- D21S1906  0.171  57.61  0.234  49.06  0.155  51.90  <<
```

```
D21S1906 -- D21S1260  0.137  59.60  0.257  46.41  0.142  64.70  >>
D21S1260 -- D21S1890  0.130  93.79  0.117 123.22  0.102 128.98  >>
D21S1890 -- D21S1885  0.020 170.35  0.031 158.07  0.022 168.63  <<
D21S1885 -- D21S1912  0.039 123.84  0.028 131.13  0.029 131.16
D21S1912 -- D21S1897  0.024 126.31  0.023 127.85  0.023 125.66  ++
D21S1897 -- D21S2057  0.093  63.33  0.075  64.30  0.034  57.47  +

Three-point distances for D21S1897:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.648  37.92  0.694   4.91  0.460  39.94  >>
 D21S263 -- D21S1252  0.463  38.55  0.396  15.06  0.449  39.17  >
D21S1252 -- D21S267   0.350  65.41  0.414  12.32  0.353  65.28
 D21S267 -- D21S1891  0.364  47.89  0.428  12.78  0.310  50.89  >>
D21S1891 -- D21S1255  0.244  94.38  0.293  26.26  0.241  94.49
D21S1255 -- D21S1893  0.293  51.87  0.267  33.15  0.250  56.92  >>
D21S1893 -- D21S266   0.218  53.30  0.203  43.27  0.179  57.55  >>
 D21S266 -- D21S1906  0.196  39.06  0.260  26.75  0.207  34.57  <<
D21S1906 -- D21S1260  0.206  41.03  0.305  24.47  0.192  45.77  >>
D21S1260 -- D21S1890  0.176  66.80  0.146  86.01  0.126  95.10  >>
D21S1890 -- D21S1885  0.033 138.03  0.042 123.42  0.035 136.61  <<
D21S1885 -- D21S1912  0.049  97.06  0.049  95.00  0.041  99.13  >>
D21S1912 -- D21S1903  0.024 126.31  0.023 125.66  0.023 127.85  >>
D21S1903 -- D21S2057  0.075  64.30  0.093  63.33  0.034  57.47  <

Three-point distances for D21S2057:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
D21S1914 -- D21S263   0.473  37.83  0.142  24.53  0.367  38.56  >
 D21S263 -- D21S1252  0.461  35.80  0.167  25.20  0.778  35.11  <
D21S1252 -- D21S267   0.512  61.31  0.086  36.31  0.474  61.41
 D21S267 -- D21S1891  0.453  43.94  0.106  29.49  0.469  43.82
D21S1891 -- D21S1255  0.503  85.05  0.063  52.13  0.511  85.03
D21S1255 -- D21S1893  0.452  42.62  0.122  34.81  0.239  45.72  >>
D21S1893 -- D21S266   0.206  40.01  0.104  34.34  0.196  40.17
 D21S266 -- D21S1906  0.177  26.83  0.094  24.39  0.175  25.87  <
D21S1906 -- D21S1260  0.208  29.82  0.080  26.59  0.206  30.29
D21S1260 -- D21S1890  0.163  46.47  0.100  45.04  0.163  48.28  >>
D21S1890 -- D21S1885  0.091  84.02  0.024  73.74  0.090  84.03
D21S1885 -- D21S1912  0.094  52.64  0.041  47.97  0.096  52.26
D21S1912 -- D21S1903  0.070  69.16  0.025  62.23  0.071  69.16
D21S1903 -- D21S1897  0.075  64.30  0.034  57.47  0.093  63.33  <

Three-point distances for D21S1914:

                           [ XX-A-B ]    [ A-XX-B ]    [ A-B-XX ]
 D21S263 -- D21S1252  0.181  79.90  0.212  68.41  0.252  58.94  <<
D21S1252 -- D21S267   0.174  82.83  0.243  43.21  0.167  83.73  >
 D21S267 -- D21S1255  0.228  69.22  0.278  34.43  0.224  70.00  >
```

```
D21S1255 -- D21S1891  0.215 102.60  0.251  38.00  0.216 102.60
D21S1891 -- D21S1893  0.285  58.43  0.380  21.50  0.331  54.47  <<
D21S1893 -- D21S1906  0.341  30.20  0.470  12.12  0.313  30.90  >
D21S1906 -- D21S266   0.331  28.50  0.609   7.72  0.386  26.55  <<
 D21S266 -- D21S1260  0.355  70.70  0.601  10.12  0.355  70.70
D21S1260 -- D21S1890  0.426  46.11  0.736   7.04  0.703  41.72  <<
D21S1890 -- D21S1885  0.592  80.09  0.859   2.81  0.578  80.18
D21S1885 -- D21S1912  0.564  50.77  0.742   3.57  0.617  50.37
D21S1912 -- D21S1903  0.597  66.35  0.894   1.96  0.613  66.27
D21S1903 -- D21S1897  0.620  59.58  1.179   0.98  0.627  59.53
D21S1897 -- D21S2057  0.734   5.67  1.022   0.73  0.658   5.69
```