NOTE TO USERS

This reproduction is the best copy available.

UMI
DETECTION OF HETEROGENEITY IN ALLELE SHARING OF AFFECTED RELATIVES

by

Lucia Mirea

A Thesis submitted in conformity with the requirements for the Degree of Master of Science
Graduate Department of Community Health
University of Toronto

© Copyright by Lucia Mirea, 2001
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

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

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

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

0-612-62931-7
Detection of Heterogeneity in Allele Sharing of Affected Relatives

Master of Science - 2001
Lucia Mirea
Graduate Department of Community Health
University of Toronto, Canada

Abstract

Linkage analyses based on allele sharing of affected relatives are less powerful when genetic heterogeneity exists among sampled families. To test for heterogeneity, we propose using family-specific non-parametric linkage (NPL) scores and family-level binary covariates that describe variability in symptoms, prevalence, and/or age of onset of the disease. First, we extend the linkage likelihood of Kong and Cox (1997) to include a family indicator covariate and formulate a likelihood ratio test and efficient score test. Simulation results show that under the null hypothesis of no heterogeneity, the empirical distributions of these test statistics do not always follow that specified by standard asymptotic theory. Alternatively, we compare mean NPL scores between two covariate-defined family subgroups using both asymptotic and permutation methods. Applied to simulations, permutation tests have nominal 5% type I error. Power decreases when the size of family subgroups is reduced or becomes unbalanced.
ACKNOWLEDGEMENTS

I am immensely grateful for the daily encouragement, patience and support of my supervisor Dr. Shelley Bull.

Many thanks to Dr. Laurent Briollais for his excellent suggestions.

I also thank my other committee members, Dr. Gerarda Darlington and Dr. Andrew Paterson, for their patience with the many delays.

Thanks to Juan Pablo Lewinger for many helpful tips.

Many thanks to Jim Wei, Larry Ng and Dennis Ng for the technical support.

I also thank Dr. Mary Corey and Len Zaifman at the Hospital for Sick Children for their assistance with the Mendel computer.

I also gratefully acknowledge all investigators of the Inflammatory Bowel Disease project.
## Contents

1 **Introduction** .......................... 1  
   1.1 Overview of Genetic Inheritance .................................. 4  
      1.1.1 Linkage and Recombination .................................. 7  
   1.2 Linkage Analysis .................................................. 8  
      1.2.1 Model-based Likelihood Analysis of Unknown Disease Loci .................................. 12  
      1.2.2 Model-free Linkage Analysis of Unknown Disease Loci .................................. 13  

2 **Linkage Analysis Based on Allele Sharing Among Sets of Affected Relatives** .......... 16  
   2.1 Extraction of Inheritance Information .................................. 17  
   2.2 Non-Parametric Linkage (NPL) Analysis .................................. 20  
   2.3 A Unified Approach to Linkage Analysis .................................. 23  
      2.3.1 An Efficient Score Test for Linkage .................................. 27  
   2.4 The Linkage Likelihood Model of Kong and Cox (1997) .................................. 28  

3 **Detection of Heterogeneity in Allele Sharing Among Sets of Affected Relatives** ...... 33  
   3.1 A Likelihood Model Allowing for Genetic Heterogeneity .................................. 35  
      3.1.1 A Likelihood Ratio Test for Heterogeneity .................................. 38  
      3.1.2 A Score Test for Heterogeneity .................................. 39
3.1.3 Consequences of Constraints ........................................ 43
3.2 Permutation Methods .................................................. 44
3.3 Tests for Heterogeneity Comparing Mean NPL Scores .......... 45

4 Simulation Study ......................................................... 48
  4.1 Overview of Simulation Study Design .......................... 49
  4.2 Simulation of Affected Pedigree Data .......................... 53
    4.2.1 Genotypes of Founding Individuals ....................... 53
    4.2.2 Construction of Nuclear and Extended Pedigree Structures . 54
    4.2.3 Genotypes for Offspring Individuals ..................... 55
    4.2.4 Phenotypes for Gender and Disease Status ............... 55
  4.3 Construction of Pedigree Samples ............................... 56
    4.3.1 Pedigree Samples Containing Locus Heterogeneity .......... 59
    4.3.2 Pedigree Samples with no Locus Heterogeneity .......... 59
    4.3.3 Data Sets with no Heterogeneity Generated with
          \( \theta = 0.25, 0.5 \) ......................................... 60
  4.4 Analysis of Simulated Data Sets ................................ 60
    4.4.1 Application of Heterogeneity Methods to Simulated Data Sets 60
    4.4.2 Methods to Summarize the Results of Heterogeneity Analyses 61
  4.5 Results of Simulation Study ...................................... 65
    4.5.1 Results of Null Data Sets with no Heterogeneity .......... 65
    4.5.2 Results of Data Sets Including Heterogeneity ........... 70
  4.6 Conclusions and Recommendations .............................. 75

5 Application of Heterogeneity Methods to Inflammatory Bowel Disease .................................................. 76
  5.1 The Data ............................................................ 78
5.2 Methods of Analyses ........................................... 80
5.3 Results .......................................................... 81
5.4 Discussion ....................................................... 83

6 Conclusions and Discussion ........................................ 89

A Derivation of Certain Results ..................................... 93
   A.1 Calculation of Maximum Likelihood Estimates for Parameters \( \delta \) and \( \beta \) 93
   A.2 Constraints for the Maximum Likelihood Estimates \( \bar{\delta}, \bar{\beta} \) .............. 95
   A.3 Results When the Parameter \( \delta \) Maximize Within the Interval \([0, b]\) 96
   A.4 Generation of Sibships \( \geq 2 \) ........................................... 100

B Results of Simulations ............................................ 102

Bibliography .......................................................... 109
# List of Figures

1.1 Meiotic products possible for two syntenic loci. ........................................ 8
1.2 A three-generation pedigree drawing with phase-known genotypes at two diallelic loci. ........................................................................................................... 9
1.3 Genotypes at a single locus for a nuclear family consisting of parents and offspring. The sibpair shares 2 alleles \textit{idb} but it is uncertain whether 1 or 2 alleles are shared \textit{idb}. ........................................ 14

2.1 Example of phase-known genotypes. ................................................................. 18

4.1 Structure of simulation design used to generate family data. ................. 51
4.2 Estimates of type I error. ................................................................................. 65
4.3 Q-Q plots comparing the observed and expected distributions of \textit{LRT}, \textit{EST} and \textit{EST}_1 methods. ................................................................. 67
4.4 Q-Q plots comparing the observed and expected distributions of \textit{LRT}, \textit{T_uv} and \textit{T_ev} methods. ................................................................. 68
4.5 Power estimates at the 5% significance level for the methods \textit{LRT}, \textit{LRT}*, \textit{EST}_{perm} and \textit{EST}_{1perm}. ................................................................. 70
4.6 Power estimates at the 5% significance level for the methods \textit{T_uV}, \textit{T_eV}, \textit{T_{uv}perm} and \textit{T_{ev}perm}. ................................................................. 71
5.1 Summary NPL scores for IBD, CD, CD16, CD> 16 and UC family subgroups ........................................ 85
5.2 Histogram of family-specific NPL scores (cM=130) .................. 86
5.3 Boxplots of family-specific NPL scores (cM=130) .................. 87
5.4 Information content for chromosome 5 .................................. 88
Chapter 1

Introduction

Following the success of existing methodology to detect genes responsible for simple genetic disorders, current research in genetic epidemiology aims to elucidate the genetic basis of complex diseases. The etiology of complex disorders is varied and may involve several susceptibility loci with interactions among multiple genetic and environmental factors. A likely component of complex disease is genetic heterogeneity which refers to the situation in which the disease trait is independently caused by two or more factors, at least one of which is genetic. Analytic approaches used in gene mapping studies, including traditional likelihood LOD scores and model-free allele-sharing methods, are generally sensitive to the underlying genetic mechanism and particularly to the presence of genetic heterogeneity. If analyses overlook the presence of genetic heterogeneity then results may be biased and conclusions misleading.

Methods that are used to localize unknown disease susceptibility genes test for linkage, or close proximity, between a suspected disease gene and marker loci at known positions throughout the genome. Conventional parametric likelihood analyses examine the inheritance pattern observed in families with multiple affected individuals, and estimate the genetic distance between a putative trait gene and each marker locus. This approach requires prior specification of a monogenic inheritance model that for complex disease is almost surely incorrect as well as difficult to specify given the wide range of possible genetic mechanisms. For this reason and also because it can be difficult to recruit large families with many affected individuals, the parametric
approach has not been used extensively in gene mapping studies of complex traits.

Alternatives to the model-based LOD score approach include model-free linkage analyses such as the transmission disequilibrium test (TDT) and allele-sharing methods. The TDT assesses preferential transmission of alleles from heterozygous parents to affected offspring. A significant result is obtained only when both linkage and allelic association are present between disease and marker loci. The TDT assumes allelic homogeneity and is best suited for precise gene localization within a linked region containing a dense set of markers.

Allele-sharing methods compare the degree of genetic similarity observed among affected relatives to that expected by chance alone. Since prior specification of a disease inheritance model is not required, allele-sharing models are often called non-parametric. Some techniques do involve parameter estimation, such as the common affected sibpair likelihood ratio test of Risch (1990b, 1990c) which requires estimation of three parameters, the respective probabilities that a sibling pair shares 0, 1 or 2 parental alleles. A method that does not require explicit estimation of parameters is the non-parametric linkage (NPL) statistic developed by Whittemore and Halpern (1994b) and implemented in the software package GENEHUNTER by Kruglyak et al. (1996). NPL analyses can be performed considering allele sharing among any pair or the entire set of affected relatives. Whittemore (1996) has demonstrated the correspondence between model-based likelihood and NPL methods, thus providing a unified approach to linkage analysis. The ensuing one-parameter likelihood model developed by Kong and Cox (1997) incorporates family-specific NPL scores, and in most situations, provides a more powerful linkage test than the NPL approach.

Both model-based and model-free methods of linkage analysis are likely to suffer a loss of statistical power when genetic heterogeneity is present within a sample of affected families. If analyses are conducted at a marker locus linked to a gene locus that contributes to disease susceptibility in only a subset of sampled families, then linkage evidence will be detected only for those families in which the linked gene locus contributes to disease susceptibility. No evidence for linkage will be observed in the remaining families where disease phenotype is due to other factors that are independent of the marker locus under consideration.
To gain a better understanding of disease etiology, it is necessary to identify the presence of genetic heterogeneity and to use methods that test for linkage in the presence of heterogeneity. Early detection of genetic heterogeneity can help researchers to focus recruitment resources toward families with a strong genetic contribution to disease susceptibility. If significant linkage findings are restricted to a particular subgroup of families, then future efforts to localize the gene can focus on this particular subgroup. Family subgroups can be identified using individual covariates, such as ethnicity, age of onset, clinical disease sub-phenotypes or the presence of an environmental factor, that may indicate the presence of genetic heterogeneity. Investigators have often grouped families according to family-level covariate classes and conducted linkage tests within such subgroups. This, however, does not provide a formal test for heterogeneity.

Tests for genetic heterogeneity have been considered in the context of both model-based and model-free approaches. Morton (1956) assessed genetic heterogeneity by comparing the recombination fraction estimates obtained from covariate-defined family subgroups. Later, Smith (1963) introduced into the parametric likelihood model a heterogeneity parameter $\alpha$ representing the proportion of families linked to the marker examined. The presence of genetic heterogeneity was determined by testing whether $\alpha < 1$. For a heterogeneous sample of affected families, the methods of Morton (1956) and Smith (1963) focus on different characteristics of heterogeneity. We distinguish between the intensity and extent of heterogeneity as investigated by Morton (1956) and Smith (1963) respectively. The intensity of heterogeneity is described by differences between the evidence of linkage observed in pedigree subgroups. In Chapter 3, we extend the likelihood model described by Kong and Cox (1997) to include a parameter $\beta$ that measures the intensity of heterogeneity between two family subgroups. As defined by Smith (1963), the parameter $\alpha$ measures the extent of heterogeneity with $\alpha = 0.5$ when heterogeneity is maximal. In the simulation study, presented in Chapter 4, $\alpha$ represents the admixture proportion of the two pedigree subgroups suspected to be heterogeneous. Based on our definition, $\alpha$ is a parameter of the covariate distribution that specifies the two pedigree subgroups. In general, our $\alpha$ represents the suspected extent of heterogeneity, but if heterogeneity is detected,
then $\alpha$ may be interpreted as the extent of heterogeneity.

Likelihood models based on allele sharing that also include covariates have been developed by Greenwood and Bull (1999) and Olson (1999a). Both of these approaches test for heterogeneity by comparing the likelihood with covariates to that without covariates. The methods of Olson (1999a) are more general and valid for any type of affected relative pair, whereas those of Greenwood and Bull (1999) are specific to affected sibling pairs. In this thesis, we develop and evaluate several tests for genetic heterogeneity that compare excess allele sharing, measured by family-level NPL scores, between two covariate-defined family subgroups. A notable difference between the proposed test statistics and those of Greenwood and Bull (1999) and Olson (1999a) is that the former are not based on pairs of affected relatives, but utilize linkage information obtained by considering allele sharing simultaneously among all affected family members.

In the remainder of this chapter we review main concepts and terminology of genetic inheritance and introduce model-based and model-free linkage methods. In Chapter 2 we provide a detailed description of the NPL approach, its correspondence to model-based likelihood analysis and the related Kong and Cox (1997) linkage model. The methods we propose to test for heterogeneity are developed in Chapter 3 and their performance is assessed via simulations described in Chapter 4. In Chapter 5, these heterogeneity tests are applied to a data set collected for a genome-wide linkage study of inflammatory bowel disease (Riou et al., 2000). We conclude with Chapter 6 providing a discussion of relevant issues and directions for future work.

1.1 Overview of Genetic Inheritance

The definitions of genetic inheritance concepts presented in this section are based on descriptions provided by Sham (1998) and Olson et al. (1999b). Genetic inheritance is the process by which deoxyribonucleic acid (DNA) is transmitted from parents to offspring. DNA has a characteristic double helix structure that is formed by a twisted string of paired molecules called nucleotide base pairs. There exist only four different types of unpaired nucleotide molecules and it is the sequence of paired nucleotides
that provides the blueprint for protein synthesis. In humans, DNA is organized into 23 different chromosome pairs located inside the nucleus of every somatic (non-gametic) cell. Normal somatic cells contain two copies of each chromosome, one inherited from the father and the other from the mother. The first 22 chromosome pairs, the *autosomes*, are homologous in size and structure. The last pair of chromosomes contains the sex chromosomes which match only in females. Females possess two X chromosomes, while males have a maternal X chromosome and one Y chromosome of paternal origin.

Genetic *locus* is a general term used to refer to any chromosomal location ranging from a single nucleotide base pair to longer DNA fragments including many consecutive base pairs. Most of the DNA molecule does not code for protein products, but contains non-coding DNA sequences placed between as well as within protein coding regions. A *gene* or *gene locus* describes a portion of DNA that includes one or more protein coding sequences whereas a *genetic marker* refers to a locus that may or may not consist of protein coding DNA. Both gene and marker loci may contain variations in the sequence of nucleotide base pairs, known as *alleles*. An individual’s *genotype* refers to the two alleles present at the same locus on homologous chromosomes. Genotypes are called *homozygous* if alleles are identical, and *heterozygous* if two different alleles are present. The pattern of alleles at loci along a single chromosome is known as a *haplotype*. Homologous haplotypes are distinguished by their parental origin.

*Phenotype* refers to the observable characteristic of a trait. Of interest in genetic studies is the relationship between a disease phenotype and the genotype present at a locus known or suspected to influence susceptibility. The probability of observing the disease phenotype given a particular genotype is called *penetrance*. Since at a locus many alleles are possible, but genotypes consist of only two alleles, penetrance probabilities must be specified for all possible two-allele combinations. Given a diallelic disease gene with disease allele *D* and normal allele *D*, three penetrance probabilities \( f_0, f_1 \) and \( f_2 \) are specified corresponding to the genotypes containing zero (\( D\bar{D} \)), one (\( D\bar{D} \) or \( \bar{D}D \)) and two (\( DD \)) disease alleles.

Assuming that disease is monogenic, inheritance can be classified as either *dominant*, *recessive* or *additive*. Inheritance is called dominant when a single disease
allele is sufficient to cause the disease phenotype and penetrances are constrained to $f_2 = f_1$ and $f_0 = 0$. When two disease alleles are necessary to develop the disease, then $f_1 = f_0 = 0$ and inheritance is considered recessive. Additive inheritance is also called codominant because the phenotype of the heterozygote is intermediate to that of the two homozygotes. Penetration probabilities for additive loci are ordered $f_2 > f_1 > f_0$ reflecting an increase in phenotype severity with an increasing number of susceptibility alleles. A gene locus is considered fully penetrant when the penetrance probabilities of susceptibility genotypes equal 1. If a disease gene is not fully penetrant there may exist individuals with susceptibility genotypes that do not exhibit the disease phenotype. It is also possible for the disease phenotype to occur in individuals that do not possess a susceptible genotype. Such disease occurrences, due to factors other than the gene locus being considered, but indistinguishable from genetic cases, are called phenocopies.

The prevalence of a monogenic trait is usually calculated assuming that the population is at Hardy-Weinberg equilibrium. This assumption allows genotype frequencies to be computed as products of population allele frequencies. If $d$ equals the frequency of the disease allele $D$, and $(1 - d)$ is the frequency of the normal allele $\bar{D}$, then the population genotype frequencies $(1 - d)^2$, $2d(1 - d)$ and $d^2$ correspond to the genotypes with zero, one and two disease alleles respectively. The prevalence of disease in the population, $K$, is then calculated according to the following relation (James, 1971).

$$K = f_2 d^2 + f_1 2d(1 - d) + f_0 (1 - d)^2 \quad (1.1)$$

The $f_0$ penetrance parameter, corresponding to the normal homozygote, represents the probability of disease caused by factors other than the gene locus under examination. These may include other genes and environmental factors that independently or jointly contribute to disease susceptibility. Inheritance is termed polygenic if development of the disease phenotype requires susceptible genotypes at two or more loci. Locus heterogeneity refers strictly to the situation when separate genes independently cause the characteristic phenotype. If both genetic and environmental factors are involved, inheritance is deemed multifactorial. The inheritance of most
complex genetic diseases is believed to be multifactorial.

1.1.1 Linkage and Recombination

Both the variability and preservation of familial traits are well explained by the transmission of DNA from parents to offspring. All human cells are diploid containing 23 chromosomes pairs except for the gametic cells involved in sexual reproduction. Parents transmit to each offspring exactly half of their DNA via gametes which are haploid and contain only a single set of chromosomes. The union of a male and female gamete forms a diploid zygote cell which has two copies of each chromosome type. The zygote cell develops into the offspring organism through mitotic cell division that preserves DNA content.

Gametes are produced by a separate cell division process called meiosis. Meiosis begins with a diploid precursor cell and produces four haploid gametes. The first step in meiosis is replication of all 23 chromosome pairs within the precursor cell thus creating four copies of each chromosome type. The cell then divides into four daughter cells, each receiving a single set of chromosomes. The placement of a chromosome type into gametes is random and independent of other chromosome types. The four gametes produced have an equal probability of forming the zygote. Consequently, the offspring has probability 1/2 of inheriting a grandpaternal or grandmaternal chromosome from each parent. This process is termed Mendel's law of independent segregation.

Due to the independent segregation of chromosomes, loci located on different chromosomes segregate independently. By the same token, syntenic loci, located on the same chromosome, have a tendency to be inherited together and are referred to as linked. Linkage is defined as the co-segregation of loci located in close proximity to each other on the same chromosome. However, not all loci positioned on the same chromosome are inherited together because during meiosis, a phenomenon called recombination or crossing-over occurs when homologous chromosomes exchange portions of DNA.

Figure 1.1 depicts the possible products of meiosis for two syntenic loci. The original non-recombinant haplotypes DM and dm are preserved if no crossing-over
or an even number of cross-overs occurs between the loci. Recombinant haplotypes $Dm$ and $dM$ are created if an odd number of cross-overs takes place. The probability of recombination $\theta$, $0 \leq \theta < 1/2$, is a function of the distance between loci. Genetic distance is measured in terms of recombination with one Morgan (M) defined as the distance in which exactly one cross-over is expected to occur. If two syntenic loci are far apart sufficient cross-overs will take place such that the loci will segregate independently corresponding to $\theta = 1/2$. When loci are linked, $\theta < 1/2$ and non-recombinant haplotypes are transmitted more often than recombinant ones.

![Parental Haplotypes](image)

![Genetic Haplotypes](image)

Figure 1.1: Meiotic products possible for two syntenic loci. Given parental haplotypes $DM$ and $dm$, an offspring will inherit either a non-recombinant ($DM, dm$) or a recombinant ($Dm, dM$) haplotype.

### 1.2 Linkage Analysis

Linkage analysis refers to the process and methods used to test whether loci are linked. The data required for linkage analysis consist of kinship relations among family members and their genotypes at the loci to be examined. The presence of linkage is indicated when co-segregation of loci within families is observed. In certain types of analyses, the relative position of two loci is determined by estimating a value for the recombination fraction $\theta$. To illustrate the process of linkage analysis, we
consider the pedigree shown in Figure 1.2 with genotypes for two diallelic loci. The pedigree schematic is drawn with males represented by squares and females by circles. At the first and second locus, the alleles possible are $A, a$ and $B, b$ respectively. For each individual, ordered genotypes are indicated as the inherited paternal/maternal haplotypes. Ordered genotypes provide information about the phase or parental origin of haplotypes and thus facilitate inference about the recombination status of transmitted haplotypes. The recombination fraction between loci is estimated as the proportion of observed recombinant haplotypes.

![Pedigree Diagram](image)

**Figure 1.2:** A three-generation pedigree drawing with phase-known genotypes at two diallelic loci.

The family in Figure 1.2 contains eight individuals, however the recombination status cannot be determined for all sixteen haplotypes. Since recombination occurs during the transmission of DNA, the haplotypes of both parents and offspring are required to infer recombination. For this reason, the distinction is made between *founders*, pedigree members from whom all others are descended and *non-founders*, individuals whose parents are included in the pedigree. In general, only the recombination status of non-founder haplotypes is considered.

There are certain situations when it is not possible to determine the recombinat-
tion status of non-founder haplotypes. For example, the recombination status of the haplotypes of individual 3, a non-founder in Figure 1.2, cannot be determined because the parents, individuals 1 and 2, are homozygous at both loci. Similarly, individual 4 is a double homozygote making it impossible to determine the recombination status of the haplotypes she has transmitted to her offspring, individuals 5, 6, 7 and 8. For this pedigree, only the paternally derived haplotypes of individuals 5, 6, 7 and 8 can be unequivocally classified as recombinant or non-recombinant. Since two of these four haplotypes are recombinants, the recombination fraction is estimated as 1/2.

For a collection of pedigrees, linkage information is combined across families and the recombination fraction is estimated as the proportion of observed recombinations $R$, given a total of $N$ non-founder haplotypes for which recombination can be determined.

$$\hat{\theta} = \frac{R}{N}$$  \hspace{1cm} (1.2)

The null hypothesis of no linkage specifies $H_0 : \theta = 1/2$. Since the recombination fraction is constrained to $0 \leq \theta \leq 1/2$, when linkage is present, the alternative one-sided hypothesis is $H_A : \theta < 1/2$. The $T$ statistic defined below (Sham, 1998) has a Chi-square distribution with 1 degree of freedom and provides a test for linkage that compares the observed number of recombinations, $R$, to that expected, $N/2$, under the null hypothesis of no linkage.

$$T = \frac{(R - \frac{N}{2})^2}{\frac{N}{2}} + \frac{(N - R - \frac{N}{2})^2}{\frac{N}{2}}$$

$$= \frac{(N - 2R)^2}{N}$$  \hspace{1cm} (1.3)

Another approach to linkage analysis is to write the probability of observed recombinations as a likelihood function with parameter $\theta$.

$$L(\theta) = \theta^R(1-\theta)^{N-R}$$  \hspace{1cm} (1.4)

Note that $\hat{\theta}$ from (1.2) is also the estimate of the recombination fraction that maximizes the above likelihood function. Linkage can be tested using the following likeli-
hood ratio test,

\[ LR = 2\ln \left( \frac{L(\hat{\theta})}{L(1/2)} \right) \]  

(1.5)

which is asymptotically a 50:50 mixture of a point probability mass at 0 and a Chi-square distribution with 1 degree of freedom (Sham, 1998). The results of linkage analyses are usually reported in terms of a LOD score, the logarithm base 10 of the likelihood ratio.

\[ LOD(t) = \log_{10} \frac{L(\hat{\theta})}{L(1/2)} \]  

(1.6)

In the above linkage analyses, haplotypes are deemed uninformative and discarded when the recombination status cannot be unequivocally determined. For most pedigree data the phase of genotypes is seldom available and haplotypes rarely provide complete information about recombination. In the presence of incomplete information, linkage analyses are conducted by conditioning on all possible sets of phase-known genotypes for the pedigree. These methods make a clear distinction between observed and ordered genotypes. Ordered genotypes are the true but often unknown data, whereas the unordered alleles observed at each locus are considered to be the locus phenotype. For each individual, the probability of the observed phenotype is calculated by conditioning on all ordered genotypes that the individual may possess. For founder individuals, the probability of ordered genotype is computed using population allele frequencies under the assumption of Hardy-Weinberg equilibrium. Probabilities of ordered genotypes for non-founders are computed conditional on the genotypes of the parents and recombination fraction \( \theta \). The resulting likelihood function \( L_i(\theta) \) describes the joint probability of observed phenotypes for all members of pedigree \( i \) (Sham 1998). The likelihood \( L(\theta) \) for \( N \) pedigrees is the product of the \( N \) pedigree specific likelihoods.

\[ L(\theta) = \prod_{i=1}^{N} L_i(\theta) \]  

(1.7)

Linkage can be tested using the likelihood ratio test \( LR \) (1.5) as described earlier.
1.2.1 Model-based Likelihood Analysis of Unknown Disease Loci

Methods of linkage analysis used to map unknown disease susceptibility genes search for chromosomal regions containing marker loci that co-segregate with the disease phenotype. These analyses assume that there exists at least one disease susceptibility gene positioned somewhere within the genome. Statistical methods are used to test for linkage between a hypothesized disease locus and a set of genetic markers spaced evenly throughout the genome at known locations.

Families with multiple affected individuals are recruited and genotyped at one or more marker loci. In this situation, both the affection status and observed marker genotypes of family members are considered phenotypic data. A simple inheritance model is specified that describes the effect of a hypothetical susceptibility locus on disease phenotype. The disease locus is assumed diallelic with disease allele frequency $d$ and penetrance probabilities $f_0$, $f_1$ and $f_2$. These parameters are used to compute for each individual the probability of the observed phenotype conditional on possible ordered genotypes. Analysis is called two-point when linkage tests (between disease and marker loci) are performed independently at each marker locus.

Multipoint linkage analyses are performed using genotype information and pre-specified inter-marker distances for a set of ordered syntenic markers. Linkage tests are conducted at variable positions along the chromosomal region assuming that this is the location of the disease gene. For every position considered, a multipoint likelihood is computed that includes information from all typed markers along the chromosome. This likelihood is compared to a similar likelihood that places the disease locus outside the defined chromosomal region, unlinked to any of the marker loci. A linkage profile is obtained by plotting the resulting LOD scores against the respective chromosomal positions. The presence of linkage is declared when $LOD \geq 3$ (Morton, 1955).

Investigators have largely opted to conduct linkage analysis of complex disease using allele-sharing methods rather than the model-based LOD score approach because the latter requires specification of an oversimplified model of disease inheritance (MOI). When the MOI is correctly specified, model-based methods are more powerful.
than allele-sharing methods (Sham, 1998). However, the statistical power of parametric LOD scores can be greatly reduced when the MOI is misspecified. Greenberg et al. (1998) have suggested analyzing complex traits by performing two parametric analyses, assuming dominant and recessive MOI's with intermediate penetrances, and retaining only the strongest result after adjusting for multiple testing. This approach was found to be only slightly less powerful than analysis using the correct MOI and more powerful than commonly used allele-sharing methods (Greenberg et al. 1998; Durner et al. 1999).

1.2.2 Model-free Linkage Analysis of Unknown Disease Loci

Model-free linkage analyses do not make any explicit assumptions about disease inheritance and do not require prior specification of an inheritance model. The premise of this approach is that relatives who display the same characteristic disease phenotype do so because they have inherited an underlying disease susceptibility genotype. Depending on the degree of relation, affected individuals are likely to have inherited the same copy of one or both disease locus haplotypes. Therefore, it is expected that affected relatives will show concordance not only at the disease locus but also at marker loci closely linked to the disease gene.

Linkage is tested by comparing, for affected relatives, the observed similarity of marker genotypes to that expected by chance alone when linkage is absent. The extent of genotype similarity is measured by the number of identical alleles present in pairs or sets of affected relatives. Many analyses use pairs of affected siblings, which at any gene locus can share 0, 1 or 2 alleles. Allele sharing is assayed as identical in state (ibs) if alleles have the same DNA sequence. When ibs alleles are copies of the same parental allele, sharing is termed identical by descent (ibd).

Sharing of ibs alleles can always be unequivocally determined if there is no missing genotype data. The same is not true for ibd sharing. For example, the two siblings shown in Figure 1.3 share 2 alleles ibs, but it is not certain whether they share 1 or 2 alleles ibd. The probability distribution of ibd allele sharing for this sibpair is 0, $\frac{1}{2}$, $\frac{1}{2}$ corresponding to 0, 1 or 2 alleles shared ibd.
Figure 1.3: Genotypes at a single locus for a nuclear family consisting of parents and offspring. The sibpair shares 2 alleles \textit{ibs} but it is uncertain whether 1 or 2 alleles are shared \textit{ibd}.

Linkage methods based on \textit{ibd} sharing are more powerful than \textit{ibs} techniques (Davies et al. 1996). The likelihood of affected sibpair data, $L(z_0, z_1, z_2)$ is written as a function of \textit{ibd} sharing parameters $z_0$, $z_1$, $z_2$ representing the probability that a sibpair shares 0, 1 and 2 alleles \textit{ibd}. Under the null hypothesis of no linkage and assuming random segregation of gametes, these take values $z_0 = \frac{1}{4}$, $z_1 = \frac{1}{2}$ and $z_2 = \frac{1}{4}$. The likelihood function is maximized with respect to the sharing parameters and linkage is tested using the likelihood ratio statistic

$$LR = 2\ln \frac{L(z_0, z_1, z_2)}{L\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)}.$$  \hspace{1cm} (1.8)

In the absence of linkage, with sharing parameters constrained to $z_0 + z_1 + z_2 = 1$, the above LR statistic has an asymptotic $\chi^2$ distribution with 2 degrees of freedom (Sham, 1998). Holmans (1993) has shown that a more powerful linkage test results if the sharing parameters are also constrained to $z_0 < \frac{1}{2}$ and $2z_0 < z_1$, as dictated by a single locus genetic mechanism.

Recently, methods have been developed that assess \textit{ibd} allele sharing for any type of relative pair (Davies et al. 1996; Curtis and Sham 1994; Kruglyak et al. 1996; Olson 1999; Cordell 2000). A popular approach is the non-parametric linkage (NPL) test (Whittemore and Halpern 1994b; Kruglyak et al. 1996) that considers allele sharing simultaneously among all affected individuals within a pedigree. A detailed
description of the NPL method is provided in the following chapter.
Chapter 2

Linkage Analysis Based on Allele Sharing Among Sets of Affected Relatives

Model-free linkage analyses of complex traits are commonly performed using the non-parametric linkage (NPL) test developed by Kruglyak et al. (1996). The NPL statistic is based on a class of linkage tests, first described by Whittemore and Halpern (1994b), that compare observed and expected \( ibd \) allele sharing among affected relatives. At any chromosomal position, an NPL score is computed separately for every family using the inheritance information extracted from the genealogy and marker genotypes of family members. Each family-specific NPL score indicates the extent to which observed \( ibd \) allele sharing among a specified set of affected relatives deviates from that expected by chance alone. Family-specific NPL scores are linearly combined to obtain a summary NPL statistic which is then used to test for linkage within the entire sample of families.

The current chapter reviews how the inheritance information is extracted from pedigree data and how this is used in NPL analysis (Kruglyak et al. 1996). We also show the correspondence between NPL and likelihood methods and describe the linkage likelihood model of Kong and Cox (1997) which is based on family-specific NPL scores.
2.1 Extraction of Inheritance Information

Considering a single marker locus at chromosomal position \( t \), a binary inheritance vector \( v(t) \) is defined that summarizes the inheritance information contained in the marker genotypes of family members. The inheritance vector \( v(t) \) indicates for each non-founding individual the outcome of the paternal and maternal meioses (Kruglyak et al. 1996). Given a family with \( nf \) non-founding individuals, the inheritance vector is

\[
v(t) = (f_1, m_1, \ldots, f_{nf}, m_{nf})
\]

where \( f_j \) (\( j = 1 \ldots nf \)) equals 0 or 1 depending on whether the \( j^{th} \) non-founder individual has inherited a grandpaternal or grandmaternal allele from the father. Similarly, \( m_j \) equals 0 or 1 depending on whether a grandpaternal or grandmaternal allele was transmitted during the maternal meiosis of individual \( j \).

Due to incomplete genotype information, it is often not feasible to determine the true inheritance vector for a family. The available genotype information is used to construct a probability distribution for all inheritance vectors that are compatible with the observed data. Given \( nf \) non-founding relatives, let \( V \) represent the set of \( 2^{2nf} \) possible inheritance vectors and let \( w \) be one realization of \( V \) such that \( w \in V \). When genotype data is completely missing, the inheritance distribution, denoted \( P_{\text{uniform}} \), is uniform over all \( 2^{2nf} \) inheritance vectors. This is the expected distribution of inheritance vectors under the null hypothesis of no linkage. As genotype data becomes available, certain inheritance vectors are excluded and probability mass becomes concentrated on those inheritance vectors remaining compatible with the observed data. The inheritance distribution conditional on the genotype data for the marker is referred to as \( P_{\text{marker}} \).

To illustrate how the inheritance distribution is computed, we consider the pedigree depicted in Figure 2.1. In this family, both founding parents are heterozygous and the phase of genotypes (paternal/maternal) is known providing complete inheritance information. The corresponding inheritance vector is \( v(t) = (0, 0, 0, 1) \). If the phase of genotypes were unknown, then it would not be possible to determine the grandparental origin of alleles transmitted to non-founders. In this case, the con-
ditional inheritance distribution $P_{\text{marker}}$ would consist of the following four equally likely inheritance vectors: $(0, 0, 0, 1)$, $(0, 1, 0, 0)$, $(1, 0, 1, 1)$ and $(1, 1, 1, 0)$. If no genotype data were available, the inheritance distribution $P_{\text{uniform}}$ would consist of the 16 equally likely inheritance vectors listed in Table 2.1.

![Diagram](image)

Figure 2.1: Phase-known genotypes with the paternally derived allele listed first. The corresponding inheritance vector is $v(t) = (0, 0, 0, 1)$.

When genotype data are incomplete, the distribution of inheritance vectors can be more accurately calculated by conditioning on the genotypes of syntenic markers. In general, the inheritance pattern across the genome is modeled as a Markov process with recombination causing transition among states (Lander and Green 1987; Kruglyak et al. 1996). The resulting multipoint conditional inheritance distribution is abbreviated $P_{\text{complete}}$ to indicate that full inheritance information has been extracted from the available data. For each marker locus the inheritance distribution, $P_{\text{marker}}$, is calculated conditional on that marker alone and considered an imperfect observation of the true inheritance pattern. For any position $t$, the conditional inheritance distribution $P_{\text{complete}}$ is constructed using the possible inheritance patterns described by $P_{\text{marker}}$ of all marker loci.

In the case of incomplete data, multipoint linkage methods are more powerful than independent analyses of marker loci because they utilize all available inheritance information. The software program GENEHUNTER developed by Kruglyak et al. (1996) can calculate both single and multipoint inheritance distributions for families where the number of founders ($f$) and that of non-founders ($nf$) conform to $2nf -$
Table 2.1: Possible inheritance vectors for the pedigree of Figure 2.1.

<table>
<thead>
<tr>
<th>Ordered Genotypes</th>
<th>Indiv. 3</th>
<th>Indiv. 4</th>
<th>$v(t)$</th>
<th>$S_{all}$</th>
<th>$P_{uniform}$</th>
<th>$P_{marker}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/C</td>
<td>A/C</td>
<td>(0,0,0,0)</td>
<td>6/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A/C</td>
<td>A/D</td>
<td>(0,0,0,1)</td>
<td>5/4</td>
<td>1/16</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>A/C</td>
<td>B/C</td>
<td>(0,0,1,0)</td>
<td>5/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A/C</td>
<td>B/D</td>
<td>(0,0,1,1)</td>
<td>1</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A/D</td>
<td>A/C</td>
<td>(0,1,0,0)</td>
<td>5/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A/D</td>
<td>A/D</td>
<td>(0,1,0,1)</td>
<td>6/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A/D</td>
<td>B/C</td>
<td>(0,1,1,0)</td>
<td>1</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A/D</td>
<td>B/D</td>
<td>(0,1,1,1)</td>
<td>5/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/C</td>
<td>A/C</td>
<td>(1,0,0,0)</td>
<td>5/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/C</td>
<td>A/D</td>
<td>(1,0,0,1)</td>
<td>1</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/C</td>
<td>B/C</td>
<td>(1,0,1,0)</td>
<td>6/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/C</td>
<td>B/D</td>
<td>(1,0,1,1)</td>
<td>5/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/D</td>
<td>A/C</td>
<td>(1,1,0,0)</td>
<td>1</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/D</td>
<td>A/D</td>
<td>(1,1,0,1)</td>
<td>5/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/D</td>
<td>B/C</td>
<td>(1,1,1,0)</td>
<td>5/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B/D</td>
<td>B/D</td>
<td>(1,1,1,1)</td>
<td>6/4</td>
<td>1/16</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
2.2 Non-Parametric Linkage (NPL) Analysis

The NPL method (Whittemore and Halpern 1994b; Kruglyak et al. 1996) tests for excess genetic similarity indicated by ibd allele sharing among affected relatives. Allele sharing is quantified by a scoring function $S(v_i(t), \Phi_i)$ that depends on the inheritance vector $v_i(t)$ and phenotype $\Phi_i$ specific to family $i$. The scoring function of each family is normalized under the null hypothesis of no linkage to obtain a family-specific NPL score. Inference about linkage is conducted using a sample summary NPL statistic computed as a linear combination of pedigree-specific NPL scores (Kruglyak et al. 1996).

A variety of scoring functions can be specified such that higher values correspond to more extensive ibd allele sharing. One possible scoring function is $S_{all}$ (Whittemore and Halpern, 1994b) which measures the degree of ibd allele sharing among the $a$ affected individuals within a family. Let $h$ represent a collection of alleles obtained by selecting one allele from each of the $a$ affected individuals, and let $b_j(h)$ equal the number of times that the $j$th founder allele appears in $h$ (for $j = 1, \ldots, 2f$). Then, $S_{all}$ is equivalent to the average number of permutations that preserves a collection obtained by choosing one allele from each affected person.

$$S_{all} = \frac{1}{2a} \sum_h \left[ \prod_{j=1}^{2f} b_j(h)! \right]$$

An alternative scoring function is $S_{pairs}$ which equals the number of alleles that two distinct affected relatives share ibd. When families include more than two affected individuals, an issue with pairwise analysis is which pairs (all or independent) to consider and how to weight these to account for possible correlation.

It has been demonstrated by McPeek (1999) that optimality of scoring functions depends more on the genetic model of inheritance at the locus under examination rather than on the degree of relation among affected individuals. Kruglyak et al. (1996) argue that the evidence for linkage is stronger when all affected relatives are...
seen to share the same allele *ibd* than when relative pairs share some alleles *ibd*. Simulation results of Kruglyak et al. (1996) showed that for dominant and complex models, NPL analysis using $S_{all}$ is more powerful than analysis using $S_{pair}$. For recessive inheritance, the performance of the two scoring functions was comparable. McPeek (1999) confirmed that $S_{all}$ is more powerful for a variety of dominant and additive models, but her results indicated $S_{pair}$ to be more powerful for the recessive case. She also cautions that the distribution of $S_{all}$ is more skewed than that of $S_{pair}$, making normal approximations less accurate.

In Chapter 3, we develop methods to test for heterogeneity in allele sharing that are based on family-level covariates and linkage scores. Both the family-level covariate and linkage score will be determined by simultaneously considering all affected individuals within a family. NPL scores will be computed using the scoring function $S_{all}$ and consequently, for the remainder of this thesis, the notation $S[\mathbf{v}(t), \Phi]$ will refer to (2.2).

As a starting point, we first consider the case of complete data for a single marker at chromosomal position $t$. Given pedigree $i$ with phenotype $\Phi_i$ and inheritance vector $\mathbf{v}_i(t)$ known with certainty, the corresponding NPL score $Z_i(t)$ is defined as

$$Z_i(t) = \frac{S[\mathbf{v}_i(t), \Phi_i] - \mu_i}{\sigma_i}$$  \hspace{1cm} (2.3)

where the terms $\mu_i$ and $\sigma_i$ are the mean and standard deviation of the scoring function calculated under the null hypothesis of no linkage. Let $P_0[\mathbf{v}_i(t) = w]$ represent the null probability of inheritance vector $w$ drawn from $P_{uniform}$ and define

$$S_{i,w}(t) = S[\mathbf{v}_i(t) = w, \Phi_i],$$  \hspace{1cm} (2.4)

$$\mu_i = E_0[S(\mathbf{v}_i(t), \Phi_i)]$$
$$= \sum_w S_{i,w}(t)P_0[\mathbf{v}_i(t) = w]$$

and

$$\sigma_i^2 = E_0[S^2(\mathbf{v}_i(t), \Phi_i)] - E_0^2[S(\mathbf{v}_i(t))].$$
For the case of complete data, the inheritance vector is known with certainty and $S[v_i(t), \Phi_i]$ can be directly calculated by enumerating all sets $h$.

Under the null hypothesis of no linkage with $\mu_i$ and $\sigma_i$ as defined, the family-specific NPL score in (2.3) has mean 0 and variance 1 (Kruglyak et al. 1996). To illustrate, we present in Table 2.1 values for $h$ and $S_{alt}$ using the pedigree data shown in Figure 2.1. The resulting NPL score in this example is 0.

For a collection of $N$ pedigrees, the overall NPL score $Z(t)$ is

$$Z(t) = \sum_{i=1}^{N} \gamma_i Z_i(t)$$

(2.6)

where $\gamma_i$ is a family-specific weight. Kruglyak et al. (1996) propose using weights $\gamma_i = \frac{1}{\sqrt{N}}$, such that under conditions of no linkage the NPL score from (2.6) follows an asymptotic standard normal distribution. Alternatively, the exact probability distribution of $Z(t)$ can be computed by considering for each of the $N$ pedigrees all possible combinations of inheritance vectors. The exact significance level is calculated as the probability of having an NPL score greater than that observed.

When full inheritance information is not available, several inheritance vectors are compatible with the observed data. The distribution of these possible inheritance vectors can be calculated using single marker information ($P_{\text{marker}}$) or multipoint methods ($P_{\text{complete}}$). For both single and multipoint calculations, the expected value of the scoring function is computed using those inheritance vectors indicated by the data. Let $g_{i,w}(t) = P[v_i(t) = w | \text{marker data}]$, where marker data refers to either a single marker in the case of single point analyses or the set of all available markers used in multipoint calculations. The expected value of the scoring function is

$$\mathbb{E}[v_i(t), \Phi] = \sum_w S_{i,w}(t) g_{i,w}(t).$$

(2.7)

Given that

$$Z_{i,w}(t) = \left[ \frac{S_{i,w}(t) - \mu_i}{\sigma_i} \right]$$

22
the expected value $\tilde{Z}_i(t)$ is the pedigree-specific NPL score.

$$\tilde{Z}_i(t) = \sum_w Z_{i,w}(t) g_{i,w}(t) \quad \text{(2.8)}$$

The $\tilde{Z}_i(t)$ are combined for the $N$ sampled families to obtain an overall NPL statistic.

$$Z(t) = \sum_{i=1}^{N} \gamma_i \tilde{Z}_i(t) = \sum_{i=1}^{N} \gamma_i \left[ \frac{\sum_w S_{i,w}(t) g_{i,w}(t) - \mu_i}{\sigma_i} \right] \quad \text{(2.9)}$$

The distribution of $Z(t)$ in (2.9) depends on the distribution of inheritance vectors from every pedigree analyzed. It can be estimated via Monte Carlo sampling from all possible realizations of the marker data. However, to assess statistical significance, Kruglyak et al. (1996) suggest using a perfect data approximation that assumes an approximate Gaussian distribution for $Z(t)$ under the null hypothesis of no linkage. The software program GENEHUNTER (Kruglyak et al. 1996) uses the perfect data approximation to calculate an exact significance level for the overall NPL statistic. Use of the perfect data approximation overestimates the variance of $Z(t)$ under the null, thus rendering NPL analysis conservative. This problem is most serious when there is very little inheritance information.

### 2.3 A Unified Approach to Linkage Analysis

A unified approach to linkage analysis was presented by Whittemore (1996) who developed a likelihood model based on $ibd$ allele sharing in general pedigrees and showed the correspondence between this model-based likelihood and NPL analysis. To construct the likelihood model of Whittemore (1996) we consider a pedigree $i$ with $M$ markers typed at positions $t_1, \ldots, t_M$ along a chromosome. Let $Y_i(t_m)$ represent the observed genotypes of the marker at position $t_m$ and let $W_i(t_m)$ denote the phase-known genotypes such that $Y_i(t_m)$ and $W_i(t_m)$ are equivalent to $ibs$ and $ibd$ allele configurations. The inheritance information contained in $W_i(t_m)$ is the same as that captured by the inheritance vector $v_i(t_m)$, and $w$ will be used to denote either a phase-known allele configuration or the corresponding inheritance vector. If a marker $m$ is
fully informative, then ibs and ibd allele configurations are the same and \( W_i(t_m) = Y_i(t_m) \). As marker information is usually incomplete, neither \( W_i(t_m) \) nor \( v_i(t_m) \) can be unequivocally determined from \( Y_i(t_m) \).

The null hypothesis of no linkage states that there is no trait predisposing gene lying within the chromosomal region spanned by the typed markers \( t_1, \ldots, t_M \). Therefore, under null conditions, the marker data and the pedigree phenotype \( \Phi_i \) are independent, and the likelihood function is equal to the probability of the observed genotypes.

\[
L_{0,i} = P_0[Y_i(t_1) = y_1, \ldots, Y_i(t_M) = y_M] \quad (2.10)
\]

The alternative hypothesis specifies that a trait gene exists at position \( t \) within the chromosomal region \( t_1, \ldots, t_M \), and that its effect on phenotype is measured by the parameter \( \delta \). Whittemore (1996) defines the pedigree-specific likelihood function as the probability of observed marker genotypes \( Y_i(t_1), \ldots, Y_i(t_M) \) conditional on the pedigree’s structure and joint phenotype \( \Phi_i \), the location of the trait gene \( t \), and \( \delta \) the effect of the gene on the trait.

\[
L_i(t, \delta) = P[Y_i(t_1) = y_1, \ldots, Y_i(t_M) = y_M|\Phi_i, t, \delta]
\]

\[
= P[\Phi_i|Y_i(t_1) = y_1, \ldots, Y_i(t_M) = y_M, t, \delta|P_0[Y_i(t_1) = y_1, \ldots, Y_i(t_M) = y_M|\Phi_i, t, \delta]](2.11)
\]

The denominator \( P(\Phi_i) \) in (2.11) is the marginal probability of observing pedigree phenotype \( \Phi_i \) independent of the marker data, and is calculated conditional on the gene effect parameter \( \delta \) and all possible ibd configurations \( w \) for the presumed trait gene.

\[
P(\Phi_i) = \sum_w P[\Phi_i|W_i(t) = w, \delta]P_{i,w}(t) \quad (2.12)
\]

where \( P_{i,w}(t) = P_0[W_i(t) = w] \) is the null probability of observing ibd configuration \( w \). Replacing \( P_0[Y_i(t_1) = y_1, \ldots, Y_i(t_M) = y_M] \) with \( L_{0,i} \) from (2.10), and conditioning on all possible ibd configurations of locus \( t \), the likelihood function in (2.11) simplifies to
\[ L_i(t, \delta) = \sum_w P[\Phi_i | W(t) = w, \delta] \frac{P[W(t) = w | Y_I(t_1) = y_1, \ldots, Y_I(t_M) = y_M]}{P[\Phi_i]} L_0,i \]

\[ = \sum_w P[\Phi_i | W(t) = w, \delta] g_{i,w}(t) L_0,i \]  

(2.13)

where as before \( g_{i,w}(t) = P[W_i(t) = w | Y_I(t_1) = y_1, \ldots, Y_I(t_M) = y_M] \) is the conditional probability of \( ibd \) configuration \( w \) given the observed marker data.

Whittemore (1996) defines the pedigree risk ratio \( R_{i,w}(t, \delta) \), as the ratio of the conditional probability of phenotype \( \Phi_i \) given \( ibd \) configuration \( w \) at locus \( t \), and the probability of observing exactly the same pedigree phenotype irrespective of the \( ibd \) configuration at \( t \) for an arbitrary pedigree of the same size and structure.

\[ R_{i,w}(t, \delta) = \frac{P[\Phi_i | W_i(t) = w, \delta]}{P[\Phi_i]} \]  

(2.14)

A simple model for the pedigree phenotype risk ratio is

\[ R_{i,w}(t, \delta) = 1 + x_{i,w}(t) \delta \]  

(2.15)

where \( x_{i,w}(t) \) is an explanatory variable dependent on the \( ibd \) configuration of the trait gene.

Incorporating the pedigree phenotype risk ratio (2.14) into (2.13), the likelihood function becomes

\[ L_i(t, \delta) = \sum_w R_{i,w}(t, \delta) g_{i,w}(t) L_0,i \]  

(2.16)

The terms \( g_{i,w}(t) \) and \( L_0,i \) in the above likelihood are calculated directly using the observed marker genotypes and population-based estimates of marker allele frequencies.

The pedigree phenotype risk ratio \( R_{i,w}(t, \delta) \) is the only component of the likelihood that depends on the unknown parameter \( \delta \) which measures the effect of the gene at \( t \) on the trait.

In the absence of linkage, the gene at \( t \) has no effect on the trait phenotype in which case \( \delta = 0 \), \( P[\Phi_i | W_i(t) = w, \delta] \equiv P[\Phi_i] \) and \( R_{i,w}(\delta_0) = 1 \). The likelihood
function under these conditions is equivalent to $L_{0,i}(t)$ defined in (2.10).

$$
L_i(t, \delta_0) = \sum_w R_{i,w}(t, 0) g_{i,w}(t) L_{0,i}
= L_{0,i} \sum_w g_{i,w}(t)
= L_{0,i}
$$

For a sample of $N$ pedigrees, the likelihood function under both the null and alternative hypotheses is obtained as the product of pedigree-specific likelihoods.

$$
L_0(t) = \prod_{i=1}^{N} L_{i,0}(t)
$$

(2.17)

$$
L(t, \delta) = \prod_{i=1}^{N} L_i(t, \delta)
= \prod_{i=1}^{N} \sum_w R_{i,w}(t, \delta) g_{i,w}(t) L_{i,0}
$$

(2.18)

To test for linkage, Whittemore (1996) specifies the null hypothesis $H_0 : \delta = 0$ and the one-sided alternative $H_A : \delta > 0$. She constructs a likelihood ratio statistic $LR(t)$ that compares the likelihood maximized with respect to $\delta$ to that under the null when $\delta_0 = 0$.

$$
LR(t) = 2ln \frac{L(t, \delta)}{L(t, \delta_0)}
= 2ln \sum_{i=1}^{N} \sum_w R_{i,w}(t, \delta) g_{i,w}(t)
$$

(2.19)

The distribution of $LR(t)$ in (2.19) is asymptotically $\chi^2$ with one degree of freedom (Whittemore, 1996). The null hypothesis of no linkage is rejected when the LOD($t$) score, defined below, exceeds a predetermined critical level.

$$
LOD(t) = log_{10} \frac{L(t, \delta)}{L_0(t)}
$$

(2.20)

$$
= \frac{LR(t)}{2ln(10)}
$$
2.3.1 An Efficient Score Test for Linkage

An alternative method to test for linkage using the likelihood function (2.18) is to use an efficient score statistic $ES(t)$. The likelihood ratio test and the efficient score method, that test the same null hypothesis, are asymptotically equivalent but the latter is computationally less demanding. Whittemore (1996) has shown that the NPL test is a special case of the efficient score ES test derived from (2.18). To calculate this $ES(t)$, let $l_i(t, \delta) = \ln[L_i(t, \delta)]$ and $l(t, \delta) = \ln[L(t, \delta)]$ denote the natural logarithm of the pedigree-specific and sample likelihood functions. The score statistic $SS(t)$ is defined as the first derivative of $l(t, \delta)$ with respect to the parameter $\delta$, evaluated under the null hypothesis when $\delta = \delta_0 = 0$.

\[
SS(t) = \left[ \frac{\partial}{\partial \delta} l(t, \delta) \right]_{\delta = \delta_0}
\]

(2.21)

\[
= \frac{\partial}{\partial \delta} \sum_{i=1}^{N} l_i(t, \delta) \bigg|_{\delta = \delta_0}
= \sum_{i=1}^{N} \frac{\partial}{\partial \delta} \ln \left\{ \sum_{w} R_{i,w}(t, \delta) g_{i,w}(t) L_{0,i}(t) \right\} \bigg|_{\delta = \delta_0}
= \sum_{i=1}^{N} \frac{L_{0,i}(t)}{\sum_{w} R_{i,w}(t, \delta) g_{i,w}(t) L_{0,i}(t)} \sum_{w} R_{i,w}(t, \delta) g_{i,w}(t) \bigg|_{\delta = \delta_0}
\]

(2.22)

Substituting $\delta_0 = 0$, makes $R_{i,w}(t, 0) = 1$ and the remaining $L_{0,i}$ in the denominator cancels with that in the numerator. The resulting score statistic is

\[
SS(t) = \sum_{i=1}^{N} \frac{\partial}{\partial \delta} \sum_{w} R_{i,w}(t, \delta) g_{i,w}(t) \bigg|_{\delta = \delta_0}
\]

(2.23)

Under the null hypothesis of no linkage,

\[
P[W_i(t) = w | Y_i(t_i) = y_1, \ldots, Y_i(t_M) = y_M] = P_0[W_i(t) = w]
\]

and

\[
g_{i,w}(t) = p_{i,w}(t).
\]
From (2.12) and (2.14) it follows that the expectation of $SS(t)$ is:

$$E_0[SS(t)] = \sum_{i=1}^{N} \frac{\partial}{\partial \delta} \sum_{w} R_{i,w}(t, \delta)p_{i,w}(t) \bigg|_{\delta = \delta_0}$$

$$= \sum_{i=1}^{N} \frac{\partial}{\partial \delta} \sum_{w} \left\{ \frac{P[\Phi_i|W_i(t) = w, \delta]}{\sum_{i,w} P[\Phi_i|W_i(t) = w, \delta]p_{i,w}(t)} \right\} p_{i,w}(t) \bigg|_{\delta = \delta_0}$$

$$= \frac{\partial}{\partial \delta} (1)$$

$$= 0.$$ 

If a trait gene does exist in the chromosomal region being analyzed, then the expectation of $SS(t)$ is non-zero. Linkage can be tested using the efficient score statistic $ES(t)$,

$$ES(t) = SS(t)V_0[SS(t)]^{-1}SS(t) \quad (2.24)$$

where $V_0[SS(t)]$ is the variance of the score statistic $SS(t)$ evaluated under the null hypothesis of no linkage. The asymptotic distribution of $ES(t)$ is $\chi^2$ with 1 degree of freedom, and significant evidence for linkage is concluded when $ES(t)$ exceeds a predefined critical value (Whittemore, 1996).

### 2.4 The Linkage Likelihood Model of Kong and Cox (1997)

Following the approach of Whittemore (1996), Kong and Cox (1997) developed a linkage likelihood model in which the covariates of the pedigree phenotype risk ratio (2.15), are weighted NPL scores. They define

$$x_{i,w} = \gamma_i Z_{i,w}(t) \quad (2.25)$$

and the corresponding pedigree phenotype risk ratio is

$$R_{i,w}(t, \delta) = [1 + \delta \gamma_i Z_{i,w}(t)]. \quad (2.26)$$

For a set of $N$ pedigrees the resulting score statistic $SS(t)$ is equivalent to the overall NPL score defined in (2.9):
When the weights \(Y_i\) are constrained to \(c::\), = 1 and the perfect data approximation of Kruglyak et al. (1996) is applied, the efficient score statistic simplifies to

\[
SS(t) = \sum_{i=1}^{N} \frac{\partial}{\partial \delta} \sum_w [1 + \delta \gamma_i Z_i(t)] g_i(t) \bigg|_{\delta = \delta_0} \\
= \sum_{i=1}^{N} \gamma_i \sum_w Z_i(t) g_i(t) \\
= Z(t).
\]

as the null variance of \(Z(t)\) is approximated by 1. As previously indicated, use of the perfect data approximation gives conservative results especially when inheritance information is only partially available. The likelihood ratio test proposed by Kong and Cox (1997) does not require the use of the perfect data approximation and provides a more accurate linkage test than \(ES(t)\) or NPL analysis for most situations where inheritance information is incomplete.

The likelihood function specified by Kong and Cox (1997) for the \(i\)th pedigree is

\[
L_i(t, \delta) = \sum_w [1 + \delta \gamma_i Z_i(t)] g_i(t) L_0,i \\
= [1 + \delta \gamma_i Z_i(t)] L_0,i. 
\]

In order to simplify notation, we will use \(Z_i(t)\) to represent the NPL score for both cases when full (2.3) or incomplete (2.8) inheritance information is available. The likelihood function for a set of \(N\) pedigrees follows as

\[
L(t, \delta) = \prod_{i=1}^{N} [1 + \delta \gamma_i Z_i(t)] L_0,i. 
\]

Testing for linkage via a likelihood ratio test requires the maximization of \(L(t, \delta)\) (2.29) with respect to \(\delta\). An upper bound \(b\) is imposed on the maximum likelihood estimate \(\hat{\delta}\) to ensure that \(R_i(t, \delta) \geq 0\). If \(a_i(t)\) is the smallest possible value that the
scoring function $S_{i,w}(t)$ can theoretically take, then $(a_i(t) - \mu_i)$ is a negative quantity and the following inequality must hold.

\[
R_i(t, \delta) \geq 0 \\
\sum_w [R_{i,w}(t, \delta) g_{i,w}(t)] \geq 0 \\
\sum_w [1 + \delta \gamma_i Z_{i,w}(t)] g_{i,w}(t) \geq 0 \\
\sum_w \left[ 1 + \delta \gamma_i \left( \frac{S_{i,w}(t) - \mu_i}{\sigma_i} \right) \right] g_{i,w}(t) \geq 0 \\
1 + \frac{(a_i(t) - \mu_i) \delta \gamma_i}{\sigma_i} \geq 0 \\
\frac{(a_i(t) - \mu_i) \delta \gamma_i}{\sigma_i} \geq -\sigma_i \\
\delta \leq \frac{[a_i(t) - \mu_i] \gamma_i}{\sigma_i} \\
\delta \leq \frac{|\mu_i - a_i(t)| \gamma_i}{\sigma_i} \\
\delta \leq \frac{b_i(t)}{\gamma_i}
\]

Specific to position $t$, $b_i(t) = \frac{\sigma_i}{|\mu_i - a_i(t)|}$ and the upper bound of $\hat{\delta}$ is $b = \min(b_i(t))$ for $i = 1 \ldots N$ sampled pedigrees. The lower bound of $\hat{\delta}$ is 0 as the model does not permit a negative gene effect. A likelihood ratio test $LR(t)$ is constructed using the likelihood (2.29) maximized under the constraint $0 \leq \hat{\delta} \leq b$.

\[
LR(t) = 2 \ln \left[ \frac{L(t, \hat{\delta})}{L(t, \delta_0)} \right] \\
= 2[l(t, \hat{\delta}) - l(t, \delta_0)]
\]  

(2.30)

Taking the natural logarithm of the likelihood function (2.29) gives

\[
l(t, \delta) = \ln \left\{ \prod_i [1 + \delta \gamma_i Z_i(t)] L_{0,i} \right\} \\
= \sum_{i=1}^{N} \ln[L_{0,i}] + \sum_{i=1}^{N} \ln[1 + \delta \gamma_i Z_i(t)] \\
= C + \sum_{i=1}^{N} \ln[1 + \delta \gamma_i Z_i(t)]
\]  

(2.31)
where \( C = \sum_{i=1}^{N} \ln[L_{0,i}] \) is a constant calculated using the observed marker data. Following the substitution of (2.31) into (2.30) the \( LR(t) \) is expressed as

\[
LR(t) = 2 \left\{ C + \sum_{i=1}^{N} \ln[1 + \hat{\delta} \gamma_i Z_i(t)] - [C + \ln(1)] \right\} \\
= 2 \sum_{i=1}^{N} \ln[1 + \hat{\delta} \gamma_i Z_i(t)].
\]

(2.32)

Equivalently one can compute the statistic

\[
Z_{lr}(t) = \sqrt{2[l(t, \hat{\delta}) - l(t, \delta_0)]} \\
= \sqrt{2 \sum_{i=1}^{N} \ln[1 + \hat{\delta} \gamma_i Z_i(t)]}
\]

(2.33)

which is well approximated by a Gaussian distribution when the number of pedigrees is large. Kong and Cox (1997) have implemented the \( Z_{lr}(t) \) linkage test in the software program GENEHUNTER-PLUS.

When inheritance information is incomplete, \( Z_{lr}(t) \) is a more powerful test than the NPL method. However, in the presence of a gene effect, the upper bound imposed on the maximum likelihood estimate of \( \delta \) restricts the amount of possible deviation. This can be a serious problem if the data set consists of a small number of pedigrees and excessive sharing is observed (Kong and Cox, 1997). If in such a situation, the normal approximation is applied to \( Z_{lr}(t) \) the significance level obtained may be too conservative. Kong and Cox (1997) briefly outline an alternative exponential model where the parameter \( \delta \) has no upper bound. The exponential model specifies

\[
P([v_i(t) = w|\delta] = p_{i,w}(t)r_i(\delta)\exp[\frac{\delta \gamma_i[S_{i,w}(t) - \mu_i]}{\sigma_i}]
\]

(2.34)

where

\[
r_i(\delta) = \left( p_{i,w}(t) \sum_w \exp\left\{\frac{\delta \gamma_i[S_{i,w}(t) - \mu_i]}{\sigma_i}\right\}\right)^{-1}
\]

is the renormalization constant necessary to ensure that \( \sum_w P[v_i(t) = w|\delta] = 1 \). Computation of \( Z_{lr}(t) \) is more demanding for the exponential model because the conditional distribution of each \( Z_i(t) \) must be calculated. In the next chapter, we will focus on the linear model and extend the likelihood function (2.29) to formulate tests to detect heterogeneity in allele sharing.
Chapter 3

Detection of Heterogeneity in Allele Sharing Among Sets of Affected Relatives

The NPL score $Z(t)$ (2.9) and the statistic $Z_{tr}(t)$ (2.33) defined by Kong and Cox (1997) are calculated by combining scores of excess $ibd$ allele sharing across the sample of affected pedigrees. Both methods assume that the same underlying genetic mechanism is active in all pedigrees sampled. The validity of this assumption is questionable for complex disease in which genetic predisposition is one of several factors contributing to disease susceptibility.

Heterogeneity among the allele-sharing distributions of sampled pedigrees can arise when genetic susceptibility varies among the sampled pedigrees. Such is the case with locus heterogeneity when two unlinked genes independently cause disease. Environmental factors can also independently affect the disease phenotype or interact with one or more genetic loci to alter the penetrance of susceptibility genotypes. Age, for example, is an important covariate that may affect the penetrance of susceptibility loci. Regardless of the cause, the presence of genetic heterogeneity impacts on the extent of observed $ibd$ allele sharing at map positions closely linked to a disease gene, with serious consequences for linkage analysis.

As an illustration we consider the presence of locus heterogeneity within a sample
of 200 nuclear families, each having a single affected sibpair and full inheritance information. Allele-sharing linkage analysis is conducted at a chromosomal position $t$ that is closely linked to a disease susceptibility locus within a subset of 100 families. We let the affected sibpair of these linked families share 2 alleles $ibd$. In the remaining 100 families there is no linkage between $t$ and a second susceptibility locus. According to Mendelian segregation, the affected sibpair of each unlinked family is expected to share 0, 1 and 2 allele $ibd$ with probability $1/4$, $1/2$ and $1/4$. The NPL score for the linked and unlinked pedigree groups is 14.14 and 0 respectively. The overall NPL score for the combined pedigree sample of 200 families is 10, considerably less than 14.14 obtained for the linked group and much less than 20, the NPL score obtained if all 200 families were linked. Although, grossly oversimplified, the present example does clearly demonstrate how the presence of locus heterogeneity can decrease the power of allele-sharing linkage methods.

In this chapter several tests for heterogeneity are developed that compare excess $ibd$ allele sharing between two predefined pedigree subgroups. These methods are not global tests for heterogeneity but are specific to locus $t$ under examination and dependent upon the criteria used to divide the pedigree sample into two subgroups. Ideally, pedigrees should be grouped such that those families segregating the trait gene through a gene linked to $t$ are separated from the rest. Grouping may be based on a binary covariate indicative of genetic heterogeneity, such as ethnicity, early vs. late disease onset and clinical sub-phenotypes of the trait. It is difficult but important to group pedigrees independently of confounding factors, such as pedigree structure and marker informativity, that will affect the ability of methods to detect $ibd$ allele sharing.

We wish to point out that these methods test for heterogeneity among the allele-sharing distribution of pedigree groups but do not provide any information about the underlying genetic mechanism of disease. The covariate used to subgroup pedigrees may give some indication as to possible causes of observed heterogeneity. We also recommend that results from these heterogeneity analyses are interpreted with great care. Significant results must consider possible confounding factors. If no significant differences are detected, then heterogeneity can be ruled out only at the locus $t$
tested and only for the covariate used to group the pedigrees. It may be the case that heterogeneity does exist, but that the covariate grouping does not separate out those families segregating the trait gene linked to the locus under study. A second possibility is that locus heterogeneity exists elsewhere in the genome, but cannot be detected at $t$ because none of the trait causing genes are linked to this locus. It may also occur that two trait genes are positioned very close to each other or lie on either side of locus $t$ such that the extent of observed linkage to locus $t$ is roughly the same for both loci. We emphasize that comparisons of allele sharing between pedigree subgroups with significant differences do suggest the presence of heterogeneity, but negative results do not necessarily rule out the possibility of heterogeneity.

3.1 A Likelihood Model Allowing for Genetic Heterogeneity

The likelihood model (2.16) developed by Whittemore (1996) is not restricted to a single parameter $\delta$, but may in general include a vector of $k$ parameters, $\delta = (\delta_1, \ldots, \delta_k)$. Additional parameters can be incorporated by extending the pedigree phenotype risk ratio,

$$R_{i,w}(t, \delta) = \left[ 1 + \sum_{j=1}^{k} \delta_j x_{i,w,j}(t) \right],$$

(3.1)

using covariates $x_{i,w,j}(t)$ specific to pedigree $i$, inheritance vector $w$ and parameter $\delta_j$. When heterogeneity is present at locus $t$, a model that assumes the same phenotype relative risk ratio $R_{i,w}(t, \delta)$ for all sampled pedigrees is incorrect. A more accurate model can be constructed by redefining $R_{i,w}(t, \delta)$ to specify higher risk for families that segregate a trait gene linked to $t$. We propose to divide the pedigree sample into two groups and specify a different $R_{i,w}(t, \delta)$ within each subgroup created. The presence of heterogeneity can be tested by comparing the degree of excess allele sharing observed within each subgroup.

To model possible genetic heterogeneity, we introduce the following two-parameter
In (3.2) \( X \) is a binary covariate, with values 1 or -1, that defines the two pedigree subgroups. When \( X_i(t) = 1 \), the pedigree phenotype risk ratio is

\[
R_{i,g1}(t, \delta, \beta) = [1 + \delta_1 \gamma_i(t)]
\]

where \( \delta_1 = \delta + \beta \). For the second group, \( X_i(t) = -1 \), \( \delta_2 = \delta - \beta \) and

\[
R_{i,g2}(t, \delta, \beta) = [1 + \delta_2 \gamma_i(t)].
\]

The subscripts \( g1 \) and \( g2 \) are used to indicate the two pedigree groups with covariate value \( X \) equal 1 and -1 respectively.

When \( \beta = 0 \), the phenotype risk ratio is the same for both pedigree groups as required under the null hypothesis of no heterogeneity. Thus, for \( H_0 : \beta = 0 \), the likelihood function for a sample of \( N \) pedigrees is equivalent to that described by Kong and Cox (1997) (2.18).

\[
L(t, \delta) = \prod_{i=1}^{N} [1 + \delta \gamma_i(t)] L_{0,i}(t)
\]

The above \( L_{0,i}(t) \) term is the likelihood of the pedigree data under the dual null hypothesis of no linkage and no heterogeneity.

The alternative hypothesis \( H_A : \beta \neq 0 \) specifies different phenotypic risk ratios for the two pedigree subgroups. For a set of \( N \) pedigrees, the likelihood function under the alternative is:

\[
L(t, \delta, \beta) = \prod_{i=1}^{N} [1 + \delta \gamma_i(t) + \beta X_i(t) \gamma_i(t)] L_{0,i}(t)
\]

The effect of gene locus \( t \) on the phenotype is measured by \( \delta_1 = \delta + \beta \) in one pedigree group and \( \delta_2 = \delta - \beta \) in the other. The parameter \( \beta \),

\[
\beta = \frac{1}{2}(\delta_1 - \delta_2)
\]
is equal to half of the difference in gene effect between the two groups. We can also write

$$\beta \gamma_i Z_i(t) = \frac{1}{2} [R_{i,g1}(t, \delta, \beta) - R_{i,g2}(t, \delta, \beta)]$$

and if $\gamma_i = 1$, then $\beta$ equals half of the difference between the group phenotype risk ratios associated with a unit change in NPL score.

The likelihood function (3.5) can be computed using maximum likelihood estimates (mle’s) $\delta$ and $\beta$. In Appendix A.1, we show that $\delta$ and $\beta$ can be evaluated as

$$\delta = \frac{\hat{\delta}_1 + \hat{\delta}_2}{2} \quad (3.6)$$

and

$$\beta = \frac{\hat{\beta}_1 - \hat{\beta}_2}{2} \quad (3.7)$$

using maximum likelihood estimates, $\hat{\delta}_1$ and $\hat{\delta}_2$. The estimates $\hat{\delta}_1$ and $\hat{\delta}_2$ have upper bounds $b_1 \geq b$ and $b_2 \geq b$ respectively, where $b = \text{min}(b_1, b_2)$ is the upper bound of $\hat{\delta}$ described in Chapter 2. We show in Appendix A.2 that the maximum likelihood estimates, $\hat{\delta}$ and $\hat{\beta}$, are also bounded by the following constraints:

$$0 \leq \hat{\delta} \leq \text{min}(b_1, b_2) = b \quad (3.8)$$

$$\hat{\delta} - b \leq \hat{\beta} \leq b - \hat{\delta} \quad (3.9)$$

To test for heterogeneity, we will define a likelihood ratio and corresponding score statistic based on the likelihood function (3.5). Note that common asymptotic results for likelihood ratio and score tests are valid only when maximum likelihood estimates are not constrained to the bounds of restricted intervals (Cox and Hinkley, 1974). In the next two sections we assume that both $\delta$ and $\beta$ parameters maximize within their allotted intervals and describe general testing methods using asymptotic results. The case when either $\hat{\delta}$ or $\hat{\beta}$ mle’s are constrained to the bounds of the parameter space is discussed in section 3.1.3.
3.1.1 A Likelihood Ratio Test for Heterogeneity

Taking the natural logarithm of the likelihood function (3.5), we obtain

\[
l(t, \delta, \beta) = \sum_{i=1}^{N} \ln \{ [1 + \delta \gamma_i Z_i(t) + \beta X_i(t) \gamma_i Z_i(t)] L_{0,i}(t) \}
\]

\[
= \sum_{i=1}^{N} \ln [L_{0,i}(t)] + \sum_{g1} \ln [1 + (\delta + \beta) \gamma_i Z_i(t)] \\
+ \sum_{g2} \ln [1 + (\delta - \beta) \gamma_i Z_i(t)] \\
= C + \sum_{g1} \ln [1 + \delta_1 \gamma_i Z_i(t)] \\
+ \sum_{g2} \ln [1 + \delta_2 \gamma_i Z_i(t)] \\
\tag{3.10}
\]

where as before \(g1\) and \(g2\) indicate the two family subgroups and \(C = \sum_{i=1}^{N} \ln [L_{0,i}(t)]\) is a constant that can be calculated from the data.

Specific to locus \(t\), the null hypothesis of no heterogeneity can be tested using the following likelihood ratio statistic.

\[
LRT(t) = 2[l(t, \tilde{\delta}, \tilde{\beta}) - l(t, \hat{\delta}, \hat{\beta}_0)] \\
\tag{3.11}
\]

In general \(\beta_0\) refers to the value of \(\beta\) under the null hypothesis. For \(LRT\) in (3.11), \(\hat{\delta}\) is the maximum likelihood estimate of \(\delta\) calculated under the null hypothesis of no heterogeneity when \(\beta_0 = 0\). The maximum likelihood estimates \(\tilde{\delta}\) and \(\tilde{\beta}\) are jointly obtained under the alternative hypothesis that allows for both linkage and heterogeneity. The likelihood ratio statistic (3.11) expands as

\[
LRT(t) = 2 \left\{ C + \sum_{i=1}^{N} \ln [1 + \tilde{\delta} \gamma_i Z_i(t) + \tilde{\beta} X_i(t) \gamma_i Z_i(t)] \\
- C + \sum_{i=1}^{N} \ln [1 + \hat{\delta} \gamma_i Z_i(t)] \right\} \\
= 2 \left\{ \sum_{g1} \ln [1 + (\tilde{\delta} + \tilde{\beta}) \gamma_i Z_i(t)] + \sum_{g2} \ln [1 + (\tilde{\delta} - \tilde{\beta}) \gamma_i Z_i(t)] \\
- \sum_{i=1}^{N} \ln [1 + \delta \gamma_i Z_i(t)] \right\}
\]
\[
\begin{align*}
= & \left\{ 2 \sum_{g_1} \ln[1 + \hat{\delta}_1 \gamma_i Z_i(t)] + 2 \sum_{g_2} \ln[1 + \hat{\delta}_2 \gamma_i Z_i(t)] \right. \\
& \left. -2 \sum_{i=1}^{N} \ln[1 + \hat{\delta} \gamma_i Z_i(t)] \right\} \\
= & \left[ Z^2_{\text{lr},g_1} + Z^2_{\text{lr},g_2} - Z^2_{\text{lr}} \right]
\end{align*}
\]

where \(Z_{\text{lr}}\) is the statistic defined in (2.33). When the maximum likelihood estimates are not constrained, the above LRT follows an asymptotic \(\chi^2\) distribution with one degree of freedom. To reject the null hypothesis of no heterogeneity, the observed LRT statistic must lie above a prespecified critical value. In the absence of heterogeneity, the probability of obtaining a test statistic greater than that observed is

\[
p\text{-value } = 2 \left[ 1 - \Phi\left(\sqrt{\text{LRT}(t)}\right) \right],
\]

where \(\Phi\) is the cumulative Gaussian distribution function.

Using weights \(\gamma_i = 1\), the quantities \(Z_{\text{lr},g_1}, Z_{\text{lr},g_2}\) and \(Z_{\text{lr}}\) can be obtained from the software package GENEHUNTER-PLUS (Kong and Cox, 1997) and the likelihood ratio statistic in (3.11) can be evaluated without direct computation of maximum likelihood estimates, \(\hat{\delta}\) and \(\hat{\beta}\).

### 3.1.2 A Score Test for Heterogeneity

A maximum likelihood estimate for the parameter \(\beta\) is not required if a score test is used to test for heterogeneity. A score test for heterogeneity based on the likelihood function (3.5) requires only one maximum likelihood estimate \(\hat{\delta}\) in contrast to the likelihood ratio approach which requires two additional maximum likelihood estimates.

Recall the notation \(\beta_0\) represents the null value of \(\beta\) and \(\hat{\delta}\) is the maximum likelihood estimate of \(\hat{\delta}\) calculated when \(\beta_0 = 0\). Giving equal weights \(\gamma_i = 1\) to each pedigree, the natural logarithm of the likelihood function (3.5) is

\[
l(t, \delta, \beta) = C + \sum_{i=1}^{N} \ln[1 + \delta Z_i(t) + \beta X_i(t)Z_i(t)]
\]

(3.13)
Let $U(t, \hat{\delta}, \beta_0)$ represent the score vector calculated under the null hypothesis of no heterogeneity.

$$U(t, \hat{\delta}, \beta_0) = \begin{pmatrix} U_1(t, \hat{\delta}, \beta_0) \\ U_2(t, \hat{\delta}, \beta_0) \end{pmatrix}$$

(3.14)

where

$$U_1(t, \hat{\delta}, \beta_0) = \left. \frac{\partial l(t, \delta, \beta)}{\partial \delta} \right|_{\delta = \hat{\delta}, \beta = \beta_0} = \sum_{i=1}^{N} \frac{Z_i(t)}{1 + \hat{\delta}Z_i(t)}$$

(3.15)

$$U_2(t, \hat{\delta}, \beta_0) = \left. \frac{\partial l(t, \delta, \beta)}{\partial \beta} \right|_{\delta = \hat{\delta}, \beta = \beta_0} = \sum_{i=1}^{N} \frac{X_i(t)Z_i(t)}{1 + \hat{\delta}Z_i(t)}$$

(3.16)

We show in Appendix 1.3 that when $\delta$ maximizes within the interval $[0, b]$ defined in Chapter 2, the expected value of the score vector $U(t, \hat{\delta}, \beta_0)$ is 0. This expectation is taken with respect to the random variable $Z_i(t)$, the family-specific NPL score, under the assumption of no heterogeneity. It follows that the asymptotic variance-covariance matrix $\Sigma_0(t, \hat{\delta}, \beta)$ of the score vector can be expressed as

$$\Sigma_0(t, \hat{\delta}, \beta) = -E_0 \begin{bmatrix} \frac{\partial^2 l(t, \delta, \beta)}{\partial \delta^2} & \frac{\partial^2 l(t, \delta, \beta)}{\partial \delta \partial \beta} \\ \frac{\partial^2 l(t, \delta, \beta)}{\partial \beta \partial \delta} & \frac{\partial^2 l(t, \delta, \beta)}{\partial \beta^2} \end{bmatrix}$$

(3.17)

To test the null hypothesis $H_0 : \beta = 0$, we construct the classical efficient score statistic $ES$ (Cox and Hinkley, 1974).

$$ES(t, \hat{\delta}, \beta_0) = U(t, \hat{\delta}, \beta_0)^T \Sigma_0^{-1}(t, \hat{\delta}, \beta_0) U(t, \hat{\delta}, \beta_0)$$

(3.18)
Commonly, $\Sigma_0(t, \hat{\delta}, \beta_0)$ is estimated using the observed Fisher information matrix $I(t, \hat{\delta}, \beta_0)$.

$$I(t, \hat{\delta}, \beta_0) = \begin{bmatrix}
\frac{\partial^2 I(t, \delta, \beta)}{\partial \delta^2} & \frac{\partial^2 I(t, \delta, \beta)}{\partial \delta \partial \beta} \\
\frac{\partial^2 I(t, \delta, \beta)}{\partial \beta \partial \delta} & \frac{\partial^2 I(t, \delta, \beta)}{\partial \beta^2}
\end{bmatrix}_{\delta = \hat{\delta}, \beta = \beta_0}$$  \hspace{1cm} (3.19)

To simplify notation, let

$$u_i(t) = \frac{Z_i(t)}{1 + \hat{\delta}Z_i(t)}$$

$$\sum_{i=1}^{N} u_i(t) = \sum_{i=1}^{N} \frac{Z_i(t)}{1 + \hat{\delta}Z_i(t)}$$

Then,

$$I(t, \hat{\delta}, \beta_0) = \begin{bmatrix} x & y \\ y & x \end{bmatrix}$$  \hspace{1cm} (3.20)

and

$$I^{-1}(t, \hat{\delta}, \beta_0) = \begin{bmatrix} a & d \\ d & a \end{bmatrix}$$  \hspace{1cm} (3.21)

where

$$x = \left[ \sum_{i=1}^{N} (u_i)^2 \right]$$  \hspace{1cm} (3.22)

$$y = \left[ \sum_{i=1}^{N} X_i(t)(u_i)^2 \right]$$  \hspace{1cm} (3.23)

$$a = \frac{x}{x^2 - y^2}$$  \hspace{1cm} (3.24)

and

$$d = \frac{-y}{x^2 - y^2}$$  \hspace{1cm} (3.25)
The resulting efficient score statistic when the variance-covariance matrix $\Sigma_0(t, \hat{\delta}, \beta_0)$ is estimated by the observed Fisher information matrix $I(t, \hat{\delta}, \beta_0)$ will be denoted $EST(t, \hat{\delta}, \beta_0)$.

$$EST(t, \hat{\delta}, \beta_0) = \left[ \begin{array}{c} U_1(t, \hat{\delta}, \beta_0) \\ U_2(t, \hat{\delta}, \beta_0) \end{array} \right]^T \left[ \begin{array}{cc} a & d \\ d & a \end{array} \right] \left[ \begin{array}{c} U_1(t, \hat{\delta}, \beta_0) \\ U_2(t, \hat{\delta}, \beta_0) \end{array} \right]$$

$$= aU_1^2(t, \hat{\delta}, \beta_0) + 2dU_1(t, \hat{\delta}, \beta_0)U_2(t, \hat{\delta}, \beta_0) + aU_2^2(t, \hat{\delta}, \beta_0)$$

$$= \frac{\left\{ \left[ \sum u_1^2(t) \right]^2 + \left[ \sum X_i(t)u_i(t) \right]^2 \right\} - 2\left[ \sum X_i(t)u_i(t) \right]\left[ \sum u_1(t) \right]\left[ \sum X_i(t)u_i(t) \right]}{4 \left[ \sum g_1 u_1^2(t) \right] \left[ \sum g_2 u_2^2(t) \right]}.$$

(3.26)

The expression for $EST(t, \hat{\delta}, \beta_0)$ can be simplified for the case when the maximum likelihood estimate (mle) $\hat{\delta}$ is not constrained. The mle $\hat{\delta}$ is obtained by setting $l'(t, \delta)$ equal to 0. Thus, when $\delta$ maximizes within the interval $[0, \beta]$, the first component of the score vector is 0. We will use $EST_1(t, \hat{\delta}, \beta_0)$ to refer to (3.26) calculated when $U_1(t, \hat{\delta}, \beta_0) = 0$.

$$EST_1(t, \hat{\delta}, \beta_0) = \left[ \begin{array}{c} 0 \\ U_2(t, \hat{\delta}, \beta_0) \end{array} \right]^T \left[ \begin{array}{cc} a & d \\ d & a \end{array} \right] \left[ \begin{array}{c} 0 \\ U_2(t, \hat{\delta}, \beta_0) \end{array} \right]$$

$$= aU_2^2(t, \hat{\delta}, \beta_0)$$

$$= \frac{\sum_{i=1}^{N} \left( \frac{Z_i(t)}{1+\delta Z_i(t)} \right)^2 \left[ \sum_{i=1}^{N} \frac{X_i(t)Z_i(t)}{1+\delta Z_i(t)} \right]^2}{\sum_{g_1} \left( \frac{Z_i(t)}{1+\delta Z_i(t)} \right)^2 \sum_{g_2} \left( \frac{Z_i(t)}{1+\delta Z_i(t)} \right)^2}.$$

(3.27)

When $\hat{\delta}$ is unconstrained the score statistics $EST(t, \hat{\delta}, \beta_0)$ and $EST_1(t, \hat{\delta}, \beta_0)$ are equivalent and expected to follow an asymptotic $\chi^2$ distribution with 1 degree of freedom. A corresponding p-value is calculated as

$$p\text{-value} = 2 \left[ 1 - \Phi \left( \sqrt{EST_1(t)} \right) \right],$$

41
where $\Phi$ is the cumulative Gaussian distribution function. The case when $\hat{\delta}$ is constrained to either 0 or $b$ is discussed in the next section.

### 3.1.3 Consequences of Constraints

If the parameter $\delta$ maximizes outside the interval $[0, b]$, the maximum likelihood estimate $\hat{\delta}$ is constrained to the nearest bound. In this situation, standard maximum likelihood results do not apply and the test statistics $LRT$ and $EST$ should not be expected to follow an asymptotic $\chi^2$ distribution. Testing can be performed assuming a $\chi^2$ distribution, but the validity of this approximation is uncertain.

Kong and Cox (1997) report conservative results for the $Z_U$ statistic when a small number of pedigrees with extensive linkage are analyzed. In the absence of heterogeneity, when linkage in both pedigree groups is substantial and $\hat{\delta}$ is constrained to upper bound $b$, we also expect the $LRT$ test to be conservative. We have proposed to compute $LRT$ (3.11) using maximum likelihood estimates $\hat{\delta}_1$ and $\hat{\delta}_2$. It may occur that only one of $\delta_1$ and $\delta_2$ subgroup estimates is constrained. This could result in a negative $LRT$ value.

When $\hat{\delta}$ is constrained to $b$, the behaviour of the score statistic $EST$ is more difficult to predict. The significance of heterogeneity test statistics $LRT$ and $EST$ can be evaluated without making any assumptions by using distribution-free methods such as permutation techniques. We note that before significance can be assessed, values for $LRT$ and $EST$ must be computed and this requires maximum likelihood estimation. As a simple alternative, we propose using a Student's $T$ statistic to test for heterogeneity between the mean NPL score of pedigree subgroups. This approach does not involve parameter estimation and permutation methods are well suited to assess significance for the two-sample problem. Theoretical and practical aspects of permutation methods are discussed in greater detail in the following section. In Section 3.3 we describe the $T_{UV}$ and $T_{EV}$ statistics for testing heterogeneity.
3.2 Permutation Methods

Permutation tests are statistical methods for hypothesis testing that provide exact tail probability values without parametric assumptions. This approach is especially well suited to problems involving two samples, such as the question regarding differences among the allele-sharing distribution of two pedigree groups. The underlying assumption of the permutation approach is that under the null hypothesis of no group differences, all units are sampled from the same distribution. In the context of genetic heterogeneity, the permutation approach assumes that in the absence of heterogeneity, the NPL scores from the two pedigree subgroups have not only the same mean, but also share a common probability distribution (Efron and Tibshirani, 1993).

For a sample of $N$ units, each with an associated measurement for the variable of interest and a dichotomous group label, let $v$ represent the vector of observed measurements and $g$ the corresponding vector of group labels. With group sizes $N_1$ and $N_2$ such that $N = N_1 + N_2$, there are $\frac{N!}{N_1!N_2!}$ possible permutations, $g^*$, of the group label vector $g$. Given a test statistic $T(v, g)$, that is a function of $v$ and $g$, a permutation distribution is constructed by calculating for each possible group label permutation $g^*$, the corresponding test statistic $T^*(v, g^*)$. Evidence against the null hypothesis is assessed by comparing the original value of $T(v, g)$ against the tail(s) probability of the permutation distribution. An exact p-value, the probability of observing a more extreme result by chance alone, is computed as the proportion of test statistics more extreme than that observed.

When sample sizes are large, the number of possible permutations can be too great to be computationally feasible and Monte Carlo methods can be used to approximate the permutation distribution of the test statistic. In practice, only a number $B > 1000$ of permutation samples are obtained from the data. To generate a permutation sample, a simple strategy is to re-sample the $N$ measurements without replacement and assign the first $N_1$ to the first group and the remaining $N_2$ to the second group. A permutation test statistic, $T^*$ is then calculated for each of the B permutation samples generated. The p-value is computed as the proportion of $T^*$ that exceed the original $T$ value.
Closely related to permutation tests, bootstrap methods use re-sampling of data to conduct non-parametric statistical inference. A bootstrap distribution is constructed by recalculating the test statistic using $B$ bootstrap samples. Bootstrap samples are generated in the same manner as permutation samples, except that sampling is performed with replacement. Due to replacement in sampling, a bootstrap sample may not preserve the original distribution of pedigree structures as does a permutation sample (Efron and Tibshirani, 1993). Also, the variance of NPL scores within a bootstrap sample will likely be different from the variance within the original pedigree sample.

### 3.3 Tests for Heterogeneity Comparing Mean NPL Scores

Provided that pedigrees are classified into two subgroups and that corresponding pedigree specific NPL scores are available, a simple way to test whether the observed linkage evidence differs between the two pedigree groups is to compare the mean NPL scores for the two groups. In the absence of heterogeneity, the NPL population means, $\mu_1(t)$ and $\mu_2(t)$, for the two pedigree classes are equal. Therefore, to test the null hypothesis of no heterogeneity, $H_0 : \mu_1(t) - \mu_2(t) = 0$, we construct the following $T_{UV}(t)$ statistic.

$$
T_{UV}(t) = \frac{[\bar{Z}_1(t) - \bar{Z}_2(t)] - [\mu_1(t) - \mu_2(t)]}{\sqrt{\frac{S_1^2(t)}{N_1} + \frac{S_2^2(t)}{N_2}}}
$$

(3.28)

In the above expression, $\bar{Z}_1(t) = \frac{1}{N_1} \sum_{i=1}^{N_1} Z_i(t)$ and $\bar{Z}_2(t) = \frac{1}{N_2} \sum_{i=1}^{N_2} Z_i(t)$ are the sample mean NPL statistics for the two pedigree groups of size $N_1$ and $N_2$. $S_1^2(t)$ and $S_2^2(t)$ are the sample variances calculated as

$$
S_1^2(t) = \frac{\sum_{i=1}^{N_1} [Z_i(t) - \bar{Z}_1(t)]^2}{(N_1 - 1)}
$$

(3.29)
\[ S^2_2(t) = \frac{\sum_{i=1}^{N_2} (Z_i(t) - \bar{Z}_2(t))^2}{N_2 - 1} \] (3.30)

The subscript (UV) indicates that the test statistic \( T_{UV}(t) \) allows the NPL scores within the two subgroups to have different variances. If we assume that the variance of NPL scores if the same within both pedigree subgroups, then a pooled sample variance \( S^2_{pooled} \) can be calculated as a weighted average of \( S^2_1 \) and \( S^2_2 \) with weights equal to the degrees of freedom corresponding to each subgroup.

\[ S^2_{pooled} = \frac{(N_1 - 1)S^2_1 + (N_2 - 1)S^2_2}{N_1 + N_2 - 2} \] (3.31)

\( S^2_{pooled} \) may then be used to estimate the common variance of NPL scores yielding \( T_{EV}(t) \), a test statistic that assumes equal variances (EV).

\[ T_{EV}(t) = \frac{[\bar{Z}_1(t) - \bar{Z}_2(t)]}{\sqrt{\frac{N_1 + N_2}{N_1 N_2}} S^2_{pooled}(t)} \] (3.32)

When linkage is present, the distributions of \( T_{UV}(t) \) and \( T_{EV}(t) \) are unknown since both test statistics are functions of the NPL score, \( Z_i(t) \). Recall that \( Z_i(t) \) is defined for every pedigree \( i \) as the expected value of a normalized scoring function, \( S_{i,w}(t) \) taken over all possible inheritance vectors \( w \in \mathcal{V} \).

\[ Z_i(t) = \sum_{w \in \mathcal{V}} \left( \frac{S_{i,w}(t) - \mu_i}{\sigma_i} \right) g_{i,w}(t) \]

The terms \( \mu_i \) and \( \sigma_i \) are the mean and standard deviation of the scoring function under the dual null hypothesis of no linkage and no heterogeneity and \( g_{i,w} \) is the probability of inheritance vector \( w \) given the genotype data. For a given pedigree structure, there are only a finite number of possible inheritance vectors \( w \), and the distributions of \( S_i(t) \) and \( Z_i(t) \) are therefore discrete. With a large number of pedigrees, we can approximation the distribution of \( \bar{Z}_1(t) \) and \( \bar{Z}_2(t) \) with a Gaussian distribution. By the Central Limit theorem, \( T_{EV}(t) \) follows an asymptotic Student’s t distribution with degrees of freedom \( df_{EV} = N_1 + N_2 - 2 \) and \( T_{UV}(t) \) has an approximate Student’s t distribution with degrees of freedom \( df_{UV} \) given below (Moore and
McCabe, 1993).

\[
df_{UV} = \frac{\left( \frac{s_{12}^2}{N_1} + \frac{s_{22}^2}{N_2} \right)^2}{\left( \frac{1}{N_1-1} \right) \left( \frac{s_{12}^2}{N_1} \right)^2 + \left( \frac{1}{N_2-1} \right) \left( \frac{s_{22}^2}{N_2} \right)^2}
\]

The test statistics \( T_{UV}(t) \) and \( T_{EV}(t) \) can also be used to test for heterogeneity without the need for distributional assumptions via resampling techniques. \( T_{UV_{perm}}(t) \) and \( T_{EV_{perm}}(t) \) refer to the test statistics in (3.28) and (3.32) where significance was assessed using permutation methods. We emphasize that the permutation approach tests for differences among the distribution of NPL scores whereas the statistics \( T_{UV}(t) \) and \( T_{EV}(t) \) test for differences between NPL groups means using asymptotic assumptions.
Chapter 4

Simulation Study

The methods developed in Chapter 3 employ family-level covariates and test for differences between subgroups of sampled pedigrees with respect to the allele-sharing distribution of affected relatives. A simulation study was conducted to evaluate and compare the performance of these heterogeneity tests. In particular, the aims of simulations included the following.

1. Assessment of distributional properties of test statistics under the null hypothesis of no heterogeneity.

   (a) Comparison between the empirical and expected distributions of the test statistics $LRT$, $EST$, $EST_1$, $T_{UV}$ and $T_{EV}$.

   (b) Estimation of type I error for tests employing asymptotic theory ($LRT$, $EST$, $EST_1$, $T_{UV}$, $T_{EV}$) and those based on permutation methods ($EST_{perm}$, $EST_{1perm}$, $T_{UVperm}$, $T_{EVperm}$). Comparison of estimated type I error rates to the nominal 5% level.

2. Assessment of power to detect heterogeneity under variable mode of disease inheritance, overall sample size and mixing proportion $\alpha$ of family subgroups.

   (a) Comparison of power for increasing $\alpha$ levels given a constant overall pedigree sample size.
(b) Comparison of power for decreased overall pedigree sample sizes given a constant proportion $\alpha$.

3. Provide recommendations regarding the use of methods with respect to nominal type I error and optimal power to detect heterogeneity. Specifically, the following questions are of interest.

(a) Use of asymptotic vs. permutation distributions.
(b) Assumption of equal variance for tests comparing differences in mean NPL scores.
(c) Comparison of likelihood-based methods ($LRT$ and $EST$ tests) vs. permutation $T$ tests.

4.1 Overview of Simulation Study Design

Our evaluation of the heterogeneity methods under consideration focuses on cases in which a genetic marker shows suggestive or significant evidence for linkage with a complex trait. Simulations were thus designed to produce replicate samples of pedigrees including multiple affected individuals with genotypes at a marker locus displaying linkage with the trait in the presence of variable levels of linkage heterogeneity. Systematic variation among the marker allele-sharing distribution of families was introduced by simulating pedigrees according to a locus heterogeneity model. Data sets of families with multiple affected individuals were simulated to include the following information: parental relations among family members, individuals' gender, disease phenotype, and genotypes at a disease gene and a polymorphic marker locus linked with recombination fraction $\theta$. The genotypes at the disease locus were used to determine the affection status of non-founding pedigree members, and then discarded. As in a real study, only marker genotypes were used in analyses.

Two types of pedigree structures were generated: nuclear families comprised of parents and two or more affected offspring, and extended three generation pedigrees with at least one pair of affected first degree cousins. Initially, genotypes at the
disease and marker loci were simulated for the two partners of the founding $F_0$ couple of every pedigree. At the disease locus, genotypes were generated according to a diallelic model with parameter $d$ indicating the population prevalence of the disease allele. Marker genotypes were constructed using $m = 10$ alleles with equal frequencies as recommended by Sham (1998) for optimal power to detect linkage. Disease and marker alleles in founding individuals were assigned independently, thus excluding the possibility of linkage disequilibrium, which is the population association of particular disease and marker alleles. Simulations proceeded with the generation of $F_1$ offspring for the founding pedigree couple $F_0$. In extended pedigrees, the $F_1$ offspring were assigned spouses and a second generation of offspring $F_2$ was obtained. The disease and marker genotypes of offspring individuals were determined by transmission of parental haplotypes. Each individual offspring received one maternal and one paternal haplotype consisting of a disease and marker allele. Haplotypes were transmitted from parent to offspring according to Mendel's law of independent segregation and a specified recombination fraction.

The affection status of non-founding individuals was determined using simple dominant and recessive inheritance models as outlined in Table 4.1. These are loosely based on models developed by Goldin and Gershon (1988) for complex psychiatric disorders with a population disease prevalence of 1%. Weeks and Harby (1995) have also used similar models in simulations examining the power to detect linkage for identical in state allele-sharing methods. Although simple and unrealistic, these models were selected as a starting point for the assessment of heterogeneity tests. Simple models were favoured in order to minimize computational complexity and the time required for simulations. For this reason, full penetrance was assigned to all susceptibility genotypes in both models. Equation (1.1) (James, 1971) was used to relate the disease prevalence $K$ to the frequency of the disease allele $d$ and penetrance probabilities $f_0$, $f_1$ and $f_2$.

The dominant model was characterized by $K = 1\%$ and penetrance constraints $f_2 = f_1$, $f_0 = 0$. In the recessive model, the population prevalence was increased to 6% to accommodate a higher disease allele frequency with full penetrance and constraints $f_1 = f_0 = 0$. Models of inheritance more appropriate for complex traits can be
developed to include incomplete penetrance, phenocopies, effects of environmental
factors and interactions between genetic and/or non-genetic components, however
generating such additional components would considerably increase the duration of
simulations.

Table 4.1: Parameter values for the population prevalence $K$, frequency of disease
allele $d$, penetrance probabilities $f_0$, $f_1$ and $f_2$ and $psib$ the proportion of sibpair
families included in simulations of dominant and recessive models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$K$</th>
<th>$d$</th>
<th>$f_0$</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$psib$</th>
</tr>
</thead>
<tbody>
<tr>
<td>dominant</td>
<td>0.01</td>
<td>0.005</td>
<td>0.005</td>
<td>1.00</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>recessive</td>
<td>0.06</td>
<td>0.25</td>
<td>0.25</td>
<td>0.00</td>
<td>1.00</td>
<td>0.65</td>
</tr>
</tbody>
</table>

According to the disease status of offspring, only nuclear families with two or
more affected siblings and extended pedigrees with at least two affected cousins were
retained for analyses. These pedigrees made up the REP-sample of each replicate. To
introduce locus heterogeneity, families were generated using one of two recombination
fractions, $\theta_1$ and $\theta_2$. The REP-sample was constructed to included a proportion $psib$
of sibpair families generated with each recombination fraction.

The parameter $psib$ is specific to each inheritance model and represents the ex-
pected proportion of affected sibpair families to be recruited when affected sibpair
and cousin pair families are sampled from the general population. The values of $psib$
used in simulations are provided in Table 4.1. These estimates were obtained for each
inheritance model as follows. A sample of 200 pedigrees of each nuclear and extended
structure were generated and only the affected sibpair and cousin pair families were
retained. From this total number of affected sibpair and cousin pair families, $psib$
was calculated as the proportion of affected sibpair families.

Maximal variability within the allele-sharing distribution was obtained by simu-
lating families using $\theta_1 = 0$ and $\theta_2 = 0.5$, the two recombination probability extrema.
The REP-sample included $G_1$ and $G_2$ families generated with $\theta_1 = 0$ and $\theta_2 = 0.5$
respectively (Table 4.3). Appropriate numbers of affected sibpair and cousin pair
families were selected from the REP-sample to create family subsets with variable sample size and covariate levels $\alpha$ (Tables 4.4 a and b). The parameter $\alpha$ refers to the binary covariate $X = 1, -1$ assigned to classify families into two groups, and represents the proportion of families with $X = 1$. In the presence of locus heterogeneity, the covariate values 1 and -1 precisely identify whether families had been generated with $\theta_1$ or $\theta_2$. Therefore, for data sets including locus heterogeneity, $\alpha$ indicates the proportion of “linked” families classified as $X = 1$. Pedigree samples, with the same covariate levels $\alpha$ but no heterogeneity, were generated by including only linked families generated with $\theta = 0$. These null $\alpha$-samples contained a covariate level $\alpha$, but this was not representative of underlying locus heterogeneity.

Figure 4.1: Structure of simulation design used to generate family data.

The structure of the simulation design is depicted in Figure 4.1. Some families from the REP-sample were included in all or most subsets created, thus introducing correlation among family subsets. Note that family subsets with and without locus heterogeneity, were constructed to contain either a fixed total number of 150 families (SAMP) or a minimum number of 75 linked families (LINK) generated with $\theta_1 = 0$. For dominant and recessive models, Table 4.2 shows the number of replications performed in the absence and presence of heterogeneity.

Some additional simulations were performed to examine the distribution of test statistics under the null hypothesis of no heterogeneity as the underlying extent of
Table 4.2: Number of replicates performed in the presence and absence of heterogeneity under dominant and recessive inheritance models.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Dominant</th>
<th>Recessive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null (no heterogeneity)</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Alternative (heterogeneity)</td>
<td>568</td>
<td>500</td>
</tr>
</tbody>
</table>

linkage was varied. These consisted of 100 replicates including samples of 150 pedigrees generated using \( \theta = 0.25 \) and \( \theta = 0.5 \) under dominant inheritance with \( \alpha = 0.8 \).

In the following sections, we describe in detail how pedigree samples were generated including assignment of genotypes to founding individuals, generation of offspring, transmission of haplotypes from parents to offspring and selection of affected sibpair and cousin pair families to create the various family subsets.

4.2 Simulation of Affected Pedigree Data

4.2.1 Genotypes of Founding Individuals

Genotype data were first simulated for the two members of the \( F_0 \) founding pedigree couple. Alleles at the disease locus were assigned conditionally requiring the presence of at least one disease allele within the genotypes of the \( F_0 \) parents. This condition prevents generation of families in which segregation of disease alleles is not possible. The required disease allele was randomly assigned to one of the four \( F_0 \) parental haplotypes as follows. A random variate from a Uniform \((0,1)\) distribution was generated, and depending upon its corresponding unit quartile, the disease allele was assigned to the first or second paternal or maternal haplotype. Disease alleles for the three remaining \( F_0 \) haplotypes were assigned with probability \( d \). In this process a separate Uniform\((0, 1)\) random variate \( u \) was generated for each haplotype and the disease allele was assigned if \( u \leq d \). Otherwise, at the disease locus, the \( F_0 \) parental haplotypes received a normal allele.
The marker locus contained \( m = 10 \) possible alleles, each assigned with equal probability \( 1/m = 0.1 \). The unit interval was divided into 10 equal sections, each section corresponding to one of the 10 possible alleles. A \( \text{Uniform}(0, 1) \) random variate was generated, and the marker allele of the indicated interval was selected. This was repeated four times, for each of the four \( F_0 \) haplotypes.

In extended pedigrees, founding individuals also included the spouses of \( F_1 \) offspring. Spousal genotypes at the disease locus were generated unconditionally with probability \( d \) of receiving a disease allele. Marker alleles were assigned to spousal haplotypes following the same process as described above.

### 4.2.2 Construction of Nuclear and Extended Pedigree Structures

The simulation of pedigrees continued with generation of \( F_1 \) offspring. Nuclear families were generated with sibship sizes of 2-5 to allow for at least one possible pair of affected siblings. Extended pedigrees always contained two \( F_1 \) offspring that formed a single branching as each was assigned a spouse and one or two \( F_2 \) offspring. The program GENEHUNTER-PLUS (Kong and Cox, 1997) used to analyze the simulated data is restricted to pedigrees where the number of founders \( f \) and non-founders \( nf \) satisfies \( (2nf - f) \leq 16 \). Large pedigrees close to this limit can be included, but analyses take considerably longer than those using pedigrees of moderate size. To avoid lengthy computations, nuclear families were generated with a maximum of 5 siblings while in each arm of extended pedigrees the largest sibship size allowed was 2.

Suarez and Van Eerdewegh (1984) have suggested modeling sibship size using random variates from a geometric distribution with parameter \( p = 0.4551 \) corresponding to a mean of 2.2 siblings per family. A random variable \( X \) from a geometric \((p)\) distribution has density function

\[
P(X = x) = p(1 - p)^{x-1} \quad x = 1, 2, 3, \ldots
\]

and is interpreted as the trial number when the first success occurs in a series of Bernoulli \((p)\) trials with probability of success \( p \). In the context of family data,
\( X \) represents sibship size. Random variates from a geometric (0.4551) distribution constrained to a maximum of 2 were used to generate sibships of size 1 or 2 for the \( F_2 \) offspring within each arm of extended pedigrees.

To obtain sibship sizes for nuclear families with at least 2 siblings, random variates must be generated from a geometric distribution left truncated at 2. As before, we could generate random variates from a geometric distribution, but this process yields single offspring that must then be discarded. In Appendix A.4 it is shown that a random variable \( X \) from a geometric distribution \((p)\) left truncated at 2 and a random variable \( Y \) from a negative binomial distribution with parameters \((1, p)\) satisfy the relation \( X = Y + 2 \). To avoid unnecessary computations, the number of offspring in nuclear families was determined by adding 2 to random variates generated from a negative binomial distribution with parameters \((1, 0.4551)\). Large values were constrained to a maximum of 5 siblings.

### 4.2.3 Genotypes for Offspring Individuals

Haplotypes were transmitted from parents to offspring according to Mendel's law of independent segregation and \( \theta \), the recombination probability between disease and marker loci. The inheritance of alleles at the disease locus was considered first. Each offspring received one paternal and one maternal allele selected with probability \( 1/2 \). Next, marker alleles were transmitted according to whether a recombination event had occurred during the parental meiosis. For each parent, having first selected the allele transmitted at the disease locus, a marker allele was selected such that the offspring inherited a recombinant haplotype with probability \( \theta \).

### 4.2.4 Phenotypes for Gender and Disease Status

Gender and affection status of family members were determined systematically for founding individuals (\( F_0 \) parents and \( F_1 \) spouses) and randomly in offspring. Always, the first \( F_0 \) parent generated was labeled male and the second female. Gender for the \( F_1 \) and \( F_2 \) offspring was assigned with probability \( 1/2 \). In extended pedigrees, the gender of founding spouses was selected to complement that of the respective \( F_1 \)
mate. The disease status of all founding individuals was labeled unknown. Disease status of offspring was determined with penetrance probability $f_0$, $f_1$ or $f_2$ depending on whether the individual's genotype contained 0, 1 or 2 disease alleles. Penetrance values specified for dominant and recessive inheritance models are given in Table 4.1. Only nuclear families with two or more affected siblings and extended pedigrees including at least one pair of affected cousins were retained for analysis.

4.3 Construction of Pedigree Samples

Pedigrees were generated only once, at the beginning of each simulation replicate. The collection of affected sibpair and cousin pair families retained for analysis within each replicate is referred to as the REP-sample and consisted of $G_1$ pedigrees generated with complete linkage ($\theta_1 = 0$) and $G_2$ pedigrees generated in the absence of linkage ($\theta_2 = 0.5$). The same proportion of sibpair families $p_{sib}$ was included in each $G_1$ and $G_2$ sets of pedigrees. For dominant and recessive models, Table 4.3 indicates the number of affected sibpair ($N_{sib} = G_1_{sib} + G_2_{sib}$) and cousin pair ($N_{cous} = G_1_{cous} + G_2_{cous}$) families contained in $G_1$ and $G_2$ of each REP-sample. The size and composition of the REP-sample were selected to include sufficient sibpair and cousin pair families within $G_1$ and $G_2$ such that pedigree sub-samples (Table 4.4) could be created containing various levels of locus heterogeneity.

Sibpair and cousin pair families were selected from the REP-sample to create samples containing a proportion $\alpha$ of families that were assigned covariate value $X = 1$. The levels of $\alpha$ examined included $\alpha = 0.5$ the most heterogenous case, and two other intermediate levels, $\alpha = 0.6$ and $\alpha = 0.8$. For each $\alpha$ level, family samples were created to contain either a fixed total sample size (SAMP) or a fixed number of linked families (LINK) selected from $G_1$. Each data sample included $n_1 = n_1_{sib} + n_1_{cous}$ and $n_2 = n_2_{sib} + n_2_{cous}$ sibpair and cousin pair families with covariate values 1 and -1 respectively. These values are listed in Tables 4.4 (a) and (b) for the SAMP and LINK $\alpha$-samples generated under dominant and recessive inheritance.

Families in group $n_1$ with covariate classification $X = 1$ were always selected from the $G_1$ families of the REP-sample. The $n_2$ group of families, assigned $X = -1$, were
Table 4.3: Number of sibpair \((G_1\text{sib}, G_2\text{sib})\) and cousin pair \((G_1\text{cous}, G_2\text{cous})\) families generated with \(\theta_1 = 0\) \((G_1 = G_1\text{sib} + G_1\text{cous})\) and \(\theta_2 = 0.5\) \((G_2 = G_2\text{sib} + G_2\text{cous})\) in the REP-sample \((N = G_1 + G_2)\) of dominant and recessive models.

<table>
<thead>
<tr>
<th>Model</th>
<th>(psib)</th>
<th>(N)</th>
<th>(N_{sib})</th>
<th>(N_{cous})</th>
<th>(\theta_1 = 0)</th>
<th>(\theta_2 = 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(G_1\text{sib})</td>
<td>(G_1\text{cous})</td>
</tr>
<tr>
<td>dominant</td>
<td>0.80</td>
<td>240</td>
<td>192</td>
<td>48</td>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>recessive</td>
<td>0.65</td>
<td>242</td>
<td>158</td>
<td>84</td>
<td>151</td>
<td>99</td>
</tr>
</tbody>
</table>

selected from \(G_1\) and \(G_2\) to create homogeneous and heterogeneous samples. Samples denoted by SAMP were constructed to include a fixed total of 150 pedigrees and variable numbers \(n_1\) and \(n_2\) of families depending on the \(\alpha\) level. In contrast, LINK samples always contained \(n_1 = 75\) families (sibpair and cousin pair) and a variable number of families in \(n_2\). The SAMP models were created to compare methods at increasing \(\alpha\) levels given a fixed sample size. For each \(\alpha\) level, corresponding SAMP and LINK samples were compared to assess the effect of decreasing the overall sample size.

SAMP and LINK criteria were selected based on the results of Goldin and Gershon (1988), such that in the presence of locus heterogeneity, linkage would be detected with 80% power at all \(\alpha\) levels. Although Goldin and Gershon (1988) studied linkage using the mean number of \(ibd\) alleles shared by affected sibpairs, they employed similar dominant and recessive models. Their results indicate that at least 50 and 30 linked sibpair families are necessary, for dominant and recessive models, to detect linkage with 80% power when \(\alpha = 0.5\). At all \(\alpha\) levels, the SAMP and LINK criteria (at least 60 linked sibpair families for the dominant, and 49 for the recessive) exceed the sibpair guidelines of Goldin and Gershon (1988) (Tables 4.4 a and b). In addition, generated samples also included \(n_1\text{cous}\) linked cousin pair families. Therefore, we expect to detect significant evidence for linkage in at least 80% of data sets generated.
Table 4.4: Samples of sibpair \( (n_1 \text{ sib}, n_2 \text{sib}) \) and cousin pair \( (n_1 \text{ cous}, n_2 \text{cous}) \) families created with sizes SAMP (a total of 150 pedigrees) and LINK (at least 75 linked families \( \theta = 0 \)) under (a) dominant inheritance, \( \psi_b = 0.8 \) and (b) recessive inheritance, \( \psi_b = 0.65 \). The covariate level \( \alpha \) measures the proportion of families with covariate value \( X = 1 \) \( (n_1 = n_1 \text{sib} + n_1 \text{cous}) \).

(a)

<table>
<thead>
<tr>
<th>Sample</th>
<th>( \alpha )</th>
<th>( n )</th>
<th>( X = 1 )</th>
<th>( X = -1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( n_1 )</td>
<td>( n_1 \text{sib} )</td>
</tr>
<tr>
<td>SAMP</td>
<td>0.8</td>
<td>150</td>
<td>120</td>
<td>96</td>
</tr>
<tr>
<td>SAMP</td>
<td>0.6</td>
<td>150</td>
<td>90</td>
<td>72</td>
</tr>
<tr>
<td>SAMP</td>
<td>0.5</td>
<td>150</td>
<td>75</td>
<td>60</td>
</tr>
<tr>
<td>LINK</td>
<td>0.8</td>
<td>94</td>
<td>75</td>
<td>60</td>
</tr>
<tr>
<td>LINK</td>
<td>0.6</td>
<td>125</td>
<td>75</td>
<td>60</td>
</tr>
<tr>
<td>LINK</td>
<td>0.5</td>
<td>150</td>
<td>75</td>
<td>60</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Sample</th>
<th>( \alpha )</th>
<th>( n )</th>
<th>( X = 1 )</th>
<th>( X = -1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( n_1 )</td>
<td>( n_1 \text{sib} )</td>
</tr>
<tr>
<td>SAMP</td>
<td>0.8</td>
<td>150</td>
<td>120</td>
<td>78</td>
</tr>
<tr>
<td>SAMP</td>
<td>0.6</td>
<td>150</td>
<td>90</td>
<td>59</td>
</tr>
<tr>
<td>SAMP</td>
<td>0.5</td>
<td>150</td>
<td>75</td>
<td>49</td>
</tr>
<tr>
<td>LINK</td>
<td>0.8</td>
<td>94</td>
<td>75</td>
<td>49</td>
</tr>
<tr>
<td>LINK</td>
<td>0.6</td>
<td>125</td>
<td>75</td>
<td>49</td>
</tr>
<tr>
<td>LINK</td>
<td>0.5</td>
<td>150</td>
<td>75</td>
<td>49</td>
</tr>
</tbody>
</table>
as described above.

4.3.1 Pedigree Samples Containing Locus Heterogeneity

Pedigree samples containing locus heterogeneity were constructed by selecting from the REP-sample appropriate numbers of sibpair and cousin pair families generated with \( \theta_1 = 0 \) and \( \theta_2 = 0.5 \). Samples included the first listed \( n_1 \text{sib}, n_1 \text{cous}, n_2 \text{sib} \) and \( n_2 \text{cous} \) families from \( G_1 \text{sib}, G_1 \text{cous}, G_2 \text{sib} \) \( G_2 \text{cous} \) respectively. The \( n_1 = n_1 \text{sib} + n_1 \text{cous} \) families selected from \( G_1 \) were assigned covariate value \( X = 1 \), while the \( n_2 = n_2 \text{sib} + n_2 \text{cous} \) families selected from \( G_2 \) were assigned covariate value \( X = -1 \). Following the above selection process, the parameter \( \alpha \) represents not only the proportion of families with \( X = 1 \) but also the proportion of linked families, i.e. the extent of heterogeneity. The values \( n_1 \text{sib}, n_1 \text{cous}, n_2 \text{sib} \) and \( n_2 \text{cous} \) listed in Tables 4.4 (a) and (b) correspond to an underlying extent of locus heterogeneity \( \alpha \), an overall proportion \( \psi_{sib} \) of sibpair families and SAMP/LINK restrictions.

4.3.2 Pedigree Samples with no Locus Heterogeneity

Pedigree samples with the same SAMP/LINK and \( \alpha \) structures were also obtained in the absence of locus heterogeneity, to examine the null distribution of the test statistics and calculate corresponding type I error rates. Since the null hypothesis of interest is that of no heterogeneity in the presence of some linkage, homogeneous data sets were constructed by including only linked pedigrees from \( G_1 \) generated with \( \theta_1 = 0 \). The number of affected sibpair (\( n_1 \text{sib}, n_2 \text{sib} \)) and cousin pair (\( n_1 \text{cous}, n_2 \text{cous} \)) families contained in these null samples are the same as those indicated in Tables 4.4 (a) and (b), but were selected from the REP-sample as follows. The first listed \( n_1 \text{sib} \) and \( n_1 \text{cous} \) families were selected from \( G_1 \text{sib} \) and \( G_1 \text{cous} \) respectively, and assigned a covariate value \( X = 1 \). Then, the next listed \( n_2 \text{sib} \) and \( n_2 \text{cous} \) families were also selected from \( G_1 \text{sib} \) and \( G_1 \text{cous} \), but assigned a covariate value \( X = -1 \). In this manner, \( \alpha \)-samples were created containing only linked pedigrees but maintaining a proportion \( \alpha \) of families with covariate value \( X = 1 \).
4.3.3 Data Sets with no Heterogeneity Generated with \( \theta = 0.25, 0.5 \)

As previously described, the bulk of simulations performed under the null hypothesis of no heterogeneity consisted of pedigrees with tight linkage (\( \theta = 0 \)). In the presence of extreme allele-sharing, the maximum likelihood estimates, of the likelihood model (3.5), used to calculate the test statistics \( LRT \) and \( EST \) are likely to be constrained to the upper bound \( b \) as defined in Chapter 3. When maximum likelihood estimates are constrained, the null distribution of the test statistics \( LRT \) and \( EST \) will deviate from that specified by asymptotic likelihood theory.

The extent of linkage present within a sample of families determines the distribution of family-specific NPL scores, and may affect the distribution of the \( T_{UV} \) and \( T_{EV} \) statistics. Additional simulations were performed to examine how the extent of linkage impacts on the distribution of test statistics \( LRT \), \( EST \), \( EST_1 \), \( T_{UV} \) and \( T_{EV} \) under the null hypothesis of no heterogeneity. Sets of 150 homogeneous families were generated under a dominant inheritance model containing either an intermediate linkage level (\( \theta = 0.25 \)) or no linkage (\( \theta = 0.5 \)). For each linkage level, 100 replicates were simulated with covariate level \( \alpha = 0.8 \), the case when sizes of pedigree subgroups are most uneven.

4.4 Analysis of Simulated Data Sets

4.4.1 Application of Heterogeneity Methods to Simulated Data Sets

Recall that the REP-sample, created at the start of every simulation replicate, consisted of affected sibpair and cousin pair families generated with \( \theta_1 = 0 \) and \( \theta_2 = 0.5 \) as outlined in Table 4.3. All pedigrees within the REP-sample were analyzed with the software program GENEHUNTER-PLUS (Kong and Cox, 1997) to obtain family-specific NPL scores. These were calculated using the scoring function \( S_{alt} \) (2.2) that considers \( ibd \) allele-sharing among all affected family members. The NPL scores from
each family were saved and used later to compute the heterogeneity test statistics.

As described in the previous section, the REP-sample was used to create subsets of families with and without heterogeneity. Families within these subsets were assigned a binary covariate $X$ taking values 1 or -1 such that a proportion $\alpha$ of families were classified as $X = 1$. The ASM linear model of GENEHUNTER-PLUS was run using all families contained in each data set to obtain estimates for the linkage parameter $\delta$ and test statistic $Z_{lr}$ (2.34). ASM analyses were repeated for each of the two covariate-defined subgroups to generate values for $\delta_1$, $Z_{lr, g1}$, $\delta_2$ and $Z_{lr, g2}$ (3.12).

The heterogeneity test statistics defined in Chapter 3 were computed using estimates $\delta$, $\delta_1$, $\delta_2$, test statistics $Z_{lr}$, $Z_{lr, g1}$ and $Z_{lr, g2}$, family-specific NPL scores and covariate $X$. The significance of all heterogeneity tests was assessed at the 5% nominal level. P-values for the methods LRT, EST, EST1, $T_{UV}$ and $T_{EV}$ were obtained as the probability of a more extreme result relative to the expected asymptotic distribution of each test statistic. The significance of $EST_{perm}$, $EST_{1, perm}$, $T_{UV, perm}$ and $T_{EV, perm}$ was evaluated by comparing each resulting test statistic to an approximate permutation distribution. The permutation distribution of $T_{UV, perm}$ and $T_{EV, perm}$ consisted of 1000 permutation samples providing p-values with three significant digits. Due to time constraints, only 100 permutation samples were obtained for $EST_{perm}$ and $EST_{1, perm}$ and consequently, the p-values for these tests include only 2 significant digits.

4.4.2 Methods to Summarize the Results of Heterogeneity Analyses

The type I error of a statistical test is defined as the probability of obtaining a significant result when the null hypothesis is true. For each heterogeneity test, the corresponding type I error was estimated as the proportion of replicates performed under the null hypothesis of no heterogeneity that yielded a significant result with a 5% critical value. The smallest type I error difference that could be detected at the 5% level was estimated to be 0.014. This estimate was obtained using a sample size calculation with variables: $p$ the estimated sample proportion, $P$ the population
proportion, \( Q = 1 - P \), and the number of simulations \( n = 1000 \) (Fleiss, 1981).

\[
|p - P| \geq z_{0.025} \sqrt{\frac{PQ}{n}} + \frac{1}{2n}
\]

(4.2)

The term \( \frac{1}{2n} \) is a continuity correction applied to achieve closer agreement between normal and binomial probabilities. To evaluate (4.2), the unknown quantity \( P \) was set to 0.05, the expected type I error rate when large sample approximations are valid.

\[
|p - P| \geq 1.96 \sqrt{\frac{(0.05)(0.95)}{1000}} + \frac{1}{2000} = 0.014
\]

To determine whether the type I error estimates were significantly different from the nominal 0.05 level, 95% confidence intervals (CI) were computed. Commonly, approximate 95% CI are calculated using when the estimated proportion satisfies \( 0.3 \leq p \leq 0.7 \). When \( p \) is close to 0 or unity, 95% CI are more accurately computed using the following lower (LCI) and upper (UCI) bounds (Fleiss, 1981).

\[
\begin{align*}
LCI &= \frac{(2np + z_{0.025}^2 - 1) - z_{0.025} \sqrt{z_{0.025}^2 + (2p(1-p) + 1)}}{2(n + z_{0.025}^2)} \\
UCI &= \frac{(2np + z_{0.025}^2 + 1) + z_{0.025} \sqrt{z_{0.025}^2 + (2p(1-p) - 1)}}{2(n + z_{0.025}^2)}
\end{align*}
\]

(4.4) (4.5)

Since we expected the type I error of test statistics to be close to 0.05, 95% CI were computed using the expressions in (4.4) and (4.5). If these 95% CI included 0.05, then the type I error rate was considered to conform to the 5% level.

The power of a statistical test is defined as the probability of a significant result when the null hypothesis is false. Power estimates are meaningful only if calculated using critical values corresponding to a given nominal type I error rate. The power of each test statistic was calculated as the proportion of replicates with a significant result at the 5% level from the total number of replicates generated in the presence of
locus heterogeneity. These power estimates are valid and can be compared only when the corresponding type I error estimates conform to the nominal 5% level. To make comparisons, power estimates for $LRT^*$, $EST^*$, $EST_1^*$, $T_{UV}^*$ and $T_{EV}^*$ tests were calculated using empirical 5% critical values. Specific to each simulation scenario under the null hypothesis of no heterogeneity, empirical 5% critical values were obtained for each method as the 95th quantile of the distribution of test statistics from replicate data sets. The expressions (4.4) and (4.5) were used to compute 95% CI for the estimated power rates.

Recall that a greater number of replicates was performed under the null hypothesis of no heterogeneity than in the presence of heterogeneity. Since the validity of the power estimate depends on the underlying type I error, we wanted to assess type I error rate with greater precision. A larger number of replicates allows for more precise confidence intervals and detection of smaller differences between the error rates of different test statistics.

Graphical methods were used to compare distributions of test statistics as well as type I error and power estimates. Q-Q plots were drawn to compare the observed and expected distribution of methods under the null hypothesis of no locus heterogeneity. The type I error and power estimates were plotted against the various $\alpha$ levels to examine differences between inheritance models and trends for SAMP and LINK sample sizes.

Nested ANOVA analyses were performed separately for SAMP and LINK samples using the logit transformation of power estimates. Factors included were disease inheritance model, group $\alpha$ level, and test statistic. The data were limited to a single observation for each combination of factor levels because, within each scenario, the entire set of replicates was used to estimate the power of heterogeneity tests. Since the data set contained no random variation, the error term in the ANOVA analysis consists of variation due to factor interactions that were not included in the model. We fit a model consisting of a main disease effect, a nested group effect within disease and a nested test effect within group. The total sum of squares $SS_T$ is written as follows:
where \( SS_{Error} \) is the residual sum of squares defined below. We assess the effect of testing method using the test statistic

\[
F = \frac{\frac{SS_{Test(\text{Group})}}{df_{Test(\text{Group})}}}{\frac{SS_{Error}}{df_{Error}}}
\]

We define the above sum of squares using the notation of Netter et al. (1985) where \( Y_{ijk} \) represents the logit of the power estimate corresponding to disease level \( i = 1 \ldots a \), group level \( j = 1 \ldots b \), and testing method \( k = 1 \ldots c \). Then,

\[
SS_T = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2
\]

\[
SS_{Disease} = bc \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2
\]

\[
SS_{Group(Disease)} = c \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{i..})^2
\]

\[
SS_{Test(\text{Group})} = a \sum_j \sum_k (\bar{Y}_{..j} - \bar{Y}_{..})^2
\]

\[
SS_{Error} = b \sum_i \sum_k (\bar{Y}_{i..k} - \bar{Y}_{i..} - \bar{Y}_{..k} + \bar{Y}_{...})^2
\]

\[+ \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{i..k} - \bar{Y}_{..j} + \bar{Y}_{i..} + \bar{Y}_{..j} + \bar{Y}_{i..k} - \bar{Y}_{...})^2\]

Initially, pairs of test statistics were compared using the above ANOVA analyses accounting for the effects of \( \alpha \) level and inheritance model. Following the result of a significant test effect, a McNemar paired analysis was performed to examine the discordance of the test statistics within each simulation scenario.

The results of simulations are presented and discussed in the next two sections.
4.5 Results of Simulation Study

Tables B1, B2, B3 and B4 in Appendix B summarize the type I error and power estimates of heterogeneity tests that were calculated using 5% critical values for each combination of inheritance model (dominant and recessive), overall sample size (SAMP, LINK) and $\alpha$ level (0.5, 0.6, 0.8). These tables also show the corresponding 95% confidence bounds that were used to determine whether the estimates of type I error conformed to the nominal 5% level.

4.5.1 Results of Null Data Sets with no Heterogeneity

In general, the type I error estimates of permutation tests, $EST_{perm}$, $EST_{1,perm}$, $T_{UV,perm}$ and $T_{EV,perm}$, did conform to the 5% level. However, under most scenarios considered, the type I error rates estimated for the test statistics $EST$, $EST_1$, $T_{UV}$ and $T_{EV}$ were large and significantly different from 0.05. The type I error estimates of permutation methods are graphically displayed in Figure 4.2 for each SAMP and LINK sample under dominant and recessive inheritance. Specific to each inheritance model, these plots show modest differences between the type I error estimates of SAMP and LINK pedigree samples. Comparisons between models indicate that type I error estimates for $T_{UV,perm}$ and $T_{EV,perm}$ increase with increasing $\alpha$ level within the recessive but not the dominant model. Figure 4.2 also includes the type I error estimates for the LRT statistic to illustrate that this test is very conservative under dominant but not recessive inheritance. Although under recessive inheritance, the LRT does become increasingly conservative ($\alpha$ levels of 0.6 and 0.8) as the sizes of pedigree subgroups become more unbalanced.

The significance of test statistics, $EST$, $EST_1$, $T_{UV}$ and $T_{EV}$, for which the type I error estimates did not conform to 0.05, was assessed relative to the expected asymptotic distribution. The empirical and expected distribution of these methods are compared in Figures 4.3 and 4.4 via Q-Q plots for pedigree samples (SAMP, $\alpha = 0.8$) generated under the null hypothesis of no heterogeneity with complete ($\theta = 0$), intermediate $\theta = 0.25$ and no linkage ($\theta = 0.5$). One observation is that the LRT heterogeneity test is gravely conservative at the two linkage extremes when $\theta = 0$ and
Figure 4.2: Type I error estimates calculated using 5% critical values. The α levels 0.5, 0.6 and 0.8 represent the proportion of families assigned a positive covariate value $X = 1$. Plots are shown for SAMP (fixed total sample size, 150 families) and LiNK (75 families with positive covariate value) replicate samples under dominant and recessive inheritance.
\( \theta = 0.5 \). This result is consistent with the type I error estimates calculated for the LRT (Figure 4.2) using pedigree samples generated with \( \theta = 0 \) and dominant inheritance. At the median linkage level, when pedigrees were simulated with \( \theta = 0.25 \), the empirical distribution of the LRT test is closer to that expected, but remains slightly conservative in the upper tail area.

The constraints imposed on maximum likelihood estimates may explain why, under the null hypothesis of no heterogeneity, the distribution of the LRT statistic deviates from expectation to a much greater extent at the two linkage extremes than in the presence of an intermediate amount of linkage. In the absence or presence of complete linkage, one or more parameter estimates used in the calculation of the LRT statistic may be constrained to the lower or upper bound. If this is the case, then asymptotic theory no longer applies to the distribution of the resulting test statistic.

In contrast to the LRT, the empirical distributions of the score tests \( EST \) and \( EST_1 \) are extremely anti-conservative in the presence of strong linkage (Figure 4.3). When linkage is absent, the distribution of \( EST \) remains anti-conservative but the departure from expectation is less pronounced than for the case of tight linkage. The empirical distribution of \( EST_1 \) in the absence of linkage is almost the same as that expected. When linkage is intermediate (\( \theta = 0.25 \)) the distributions of \( EST \) and \( EST_1 \) are identical and appear to follow the expected \( \chi^2 \) distribution.

Since both LRT and EST are based on the likelihood model (3.5), they are asymptotically equivalent, and the presence of constrained maximum likelihood estimates will affect the distribution of both. Recall, that the \( EST_1 \) test statistic is computed only when the linkage parameter \( \delta \) maximizes within the imposed lower and upper bounds. Therefore, any observed differences between the distributions of \( EST \) and \( EST_1 \) are most likely due to constrained \( \delta \) estimates. Note that, as expected, at the two linkage extremes, a greater departure from expectation is seen for the \( EST \) than the \( EST_1 \) test. Also note that although the \( EST_1 \) statistic is computed using unconstrained maximum likelihood estimates, in the presence of tight linkage, a large deviation is observed between the empirical and expected distributions of \( EST_1 \). This indicates that the application of constraints is only one factor affecting the distribution of the likelihood-based tests LRT and EST.
Figure 4.3: Q-Q plots comparing the observed and expected distributions of LRT, EST and EST1 test statistics. Data sets contain 150 pedigrees generated under the null hypothesis of no heterogeneity with dominant inheritance. Pedigree group sizes correspond to $\alpha = 0.8$ which is the proportion of pedigrees classified as $X = 1$. Theta is the recombination fraction between the disease and marker loci and $n$ is the number of replicates analyzed.
Figure 4.4: Q-Q plots comparing the observed and expected distributions of LRT, $T_{UV}$ and $T_{EV}$ statistics. Data sets contain 150 pedigrees generated under the null hypothesis of no heterogeneity with dominant inheritance. Pedigree group sizes correspond to $\alpha = 0.8$ which is the proportion of pedigrees classified as $X = 1$. Theta is the recombination fraction between disease and marker loci and $n$ is the number of replicates analyzed.
The Q-Q plots for $T_{UV}$ and $T_{EV}$ in Figure 4.4 show good correspondance between empirical and expected distributions when $\theta = 0$ and $\theta = 0.25$. However, the type I error estimates (Tables B1) for $T_{UV}$ and $T_{EV}$ in the presence of complete linkage $\theta = 0$ (Figure 4.4) are significantly lower than 0.05, indicating that tests are conservative. When linkage is not present ($\theta = 0.5$), the correspondance between the empirical and expected distribution is less precise, for both $T_{UV}$ and $T_{EV}$, especially within the tails of the distribution.

4.5.2 Results of Data Sets Including Heterogeneity

The power estimates reported in Appendix tables B.2 and B.4 were calculated using critical values at the nominal 5% level and thus are valid only for methods with a type I error of 5%. Subsequently, we discuss only the power estimates of permutation methods, for which the type I error estimates conformed to the 5% nominal level. We will also consider the power of $LRT^*$, $EST^*$, $T_{UV}$ and $T_{EV}$ tests that was calculated using empirical 5% critical values. Figures 4.5 and 4.6 show power curves that in general decrease as the size of pedigree subgroups becomes less balanced. The observed decrease in power is more severe for the recessive model and the smaller LINK samples. The power estimates of $EST^*$ and $EST_1^*$ are not shown as these lie below 0.6 in all simulation scenarios considered.

The power estimates of permutation score tests $EST_{perm}$ and $EST_{1perm}$ are very similar to each other, but lower than the those of the $LRT^*$ (Figure 4.5). Among permutation methods, highest power rates are observed for $T_{EVperm}$ under all scenarios analyzed. In general, the power of $T_{EVperm}$ is comparable to that of $LRT^*$, $T_{UV}$ and $T_{EV}$ (Figure 4.6).

Nested ANOVA models (Table 4.5) were fit separately for SAMP and LINK pedigree samples to test whether variation among the logit of power estimates can be explained by simulation factors and methods of testing heterogeneity. The factors considered were: disease model with levels dominant and recessive, a group factor with $\alpha$ levels 0.5, 0.6 and 0.8, and a factor for the test statistics compared. A first model including test levels $EST_{perm}$, $T_{UVperm}$ and $T_{EVperm}$ showed a significant test
Figure 4.5: Power estimates at the 5% significance level for the methods $LRT$, $LRT^*$, $EST_{perm}$ and $EST_{1perm}$. Pedigree samples contain a proportion $\alpha$ of linked ($\theta = 0$) families identified by covariate value $X = 1$. SAMP and LINK groups correspond to a fixed sample size of 150 pedigrees and a minimum of 75 linked families respectively. Power estimates for $LRT^*$ were calculated using empirical 5% critical values.
Figure 4.6: Power estimates at the 5% significance level for the methods $T^*_U$, $T^*_E$, $T^*_{UV_{perm}}$ and $T^*_{EV_{perm}}$. Pedigree samples contain a proportion $\alpha$ of linked ($\theta = 0$) families identified by covariate value $X = 1$. SAMP and LINK groups correspond to a fixed sample size of 150 pedigrees and a minimum of 75 linked families. Power estimates for $T^*_{UV}$ and $T^*_{EV}$ were calculated using empirical 5% critical values.
effect while accounting for the effects of inheritance model and $\alpha$ level. In a second model, no significant difference was observed between the $T_U V_{perm}$ and $T_E V_{perm}$ test statistics, but a significant effect was detected in a third model comparing $E S T_{perm}$ and $T_E V_{perm}$ methods. Among $T_U V_{perm}$ and $T_E V_{perm}$, the latter was selected because its power estimates tended to be higher (Figure 4.6). McNemar 1 df paired comparisons (Table 4.6) showed that differences in agreement between the tests statistics $E S T_{perm}$ and $T_E V_{perm}$ were significant only at $\alpha = 0.8$. This result further emphasizes the effect of subgroup sample size on the power to detect heterogeneity.

Table 4.5: Nested ANOVA models fit for SAMP and LINK pedigree samples examining variation of the logit of power estimates. Factors included disease model (dominant and recessive), $\alpha$ level (0.5, 0.6, 0.8) and test statistic($E S T_{perm}$, $T_U V_{perm}$ and $T_E V_{perm}$). F values for model effects correspond to type I sum of squares.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>df</th>
<th>SAMP</th>
<th>LNK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td>pval</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>11</td>
<td>21.63</td>
<td>0.0006</td>
</tr>
<tr>
<td>Disease</td>
<td>dom, rec</td>
<td>1</td>
<td>106.31</td>
<td>0.0001</td>
</tr>
<tr>
<td>Group(Disease)</td>
<td>0.5, 0.6, 0.8</td>
<td>4</td>
<td>18.45</td>
<td>0.0016</td>
</tr>
<tr>
<td>Test(Group)</td>
<td>ESTperm, Tuvperm, Tevperm</td>
<td>6</td>
<td>9.63</td>
<td>0.0072</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>8</td>
<td>139.40</td>
<td>0.0009</td>
</tr>
<tr>
<td>Disease</td>
<td>dom, rec</td>
<td>1</td>
<td>846.11</td>
<td>0.0001</td>
</tr>
<tr>
<td>Group(Disease)</td>
<td>0.5, 0.6, 0.8</td>
<td>4</td>
<td>66.11</td>
<td>0.0030</td>
</tr>
<tr>
<td>Test(Group)</td>
<td>Tuvperm, Tevperm</td>
<td>3</td>
<td>1.55</td>
<td>0.3640</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>8</td>
<td>13.29</td>
<td>0.0284</td>
</tr>
<tr>
<td>Disease</td>
<td>dom, rec</td>
<td>1</td>
<td>41.00</td>
<td>0.0077</td>
</tr>
<tr>
<td>Group(Disease)</td>
<td>0.5, 0.6, 0.8</td>
<td>4</td>
<td>9.26</td>
<td>0.0490</td>
</tr>
<tr>
<td>Test(Group)</td>
<td>ESTperm, Tevperm</td>
<td>3</td>
<td>9.44</td>
<td>0.0488</td>
</tr>
</tbody>
</table>
Table 4.6: McNemar paired analysis comparing agreement between $EST_{perm}$ and $TEV_{perm}$ heterogeneity tests. $N_{reps}$ is the number of replicates, $N_{12}$ represents the counts where $EST_{perm}$ is significant but $TEV_{perm}$ is not. $N_{21}$ is the number of replicates where the opposite is true, $TEV_{perm}$ is significant but $EST_{perm}$ is not. Replications are considered for dominant (DOM) and recessive (REC) inheritance models, SAMP (150 families) and LINK (at least 75 linked families) sample sizes and $\alpha$ levels representing the proportion of families with covariate value $X = 1$.

<table>
<thead>
<tr>
<th>Disease Model</th>
<th>Sample</th>
<th>$\alpha$</th>
<th>$N_{reps}$</th>
<th>$N_{12}$</th>
<th>$N_{21}$</th>
<th>McNemar</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOM</td>
<td>SAMP</td>
<td>0.5</td>
<td>568</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DOM</td>
<td>SAMP</td>
<td>0.6</td>
<td>568</td>
<td>0</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DOM</td>
<td>SAMP</td>
<td>0.8</td>
<td>568</td>
<td>1</td>
<td>11</td>
<td>8.333</td>
<td>0.004</td>
</tr>
<tr>
<td>DOM</td>
<td>LINK</td>
<td>0.5</td>
<td>568</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DOM</td>
<td>LINK</td>
<td>0.6</td>
<td>568</td>
<td>0</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DOM</td>
<td>LINK</td>
<td>0.8</td>
<td>568</td>
<td>6</td>
<td>19</td>
<td>6.760</td>
<td>0.009</td>
</tr>
<tr>
<td>REC</td>
<td>SAMP</td>
<td>0.5</td>
<td>500</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>REC</td>
<td>SAMP</td>
<td>0.6</td>
<td>500</td>
<td>2</td>
<td>3</td>
<td>0.200</td>
<td>0.655</td>
</tr>
<tr>
<td>REC</td>
<td>SAMP</td>
<td>0.8</td>
<td>500</td>
<td>6</td>
<td>60</td>
<td>44.182</td>
<td>0.001</td>
</tr>
<tr>
<td>REC</td>
<td>LINK</td>
<td>0.5</td>
<td>500</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>REC</td>
<td>LINK</td>
<td>0.6</td>
<td>500</td>
<td>4</td>
<td>7</td>
<td>0.818</td>
<td>0.366</td>
</tr>
<tr>
<td>REC</td>
<td>LINK</td>
<td>0.8</td>
<td>500</td>
<td>16</td>
<td>93</td>
<td>54.394</td>
<td>0.001</td>
</tr>
</tbody>
</table>
4.6 Conclusions and Recommendations

The results of the simulation study clearly indicate that under the null hypothesis of no heterogeneity, the empirical distribution of test statistics LRT, EST, EST₁, TUV and TEV depends upon the extent of linkage present within the pedigree sample analyzed. Because the empirical distribution of test statistics does not always agree with large sample approximations, asymptotic theory cannot be generally applied to assess the significance of heterogeneity methods in situations when the extent of linkage is unknown. Therefore, we recommend the use of permutation tests, which for the simulated scenarios examined, had nominal 5% type I error. In general, the power estimates of permutation methods were lower for the recessive than the dominant model. Power estimates also decreased when the overall number of families was reduced or the sizes of family subgroups became less even. Although no significant difference was detected between the power estimates of TUVperm and TEVperm, the power estimates of TEVperm were greater. From among heterogeneity methods examined, we suggest the use of TEVperm because it does not require maximum likelihood estimation and under all scenarios investigated, its power is at least as great as that of the other permutation tests (TUVperm, ESTperm and EST₁perm).
Chapter 5

Application of Heterogeneity Methods to Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is a disorder of the autoimmune system characterized by chronic inflammation of the gastrointestinal (GI) tract. The prevalence of IBD is estimated at 1-2/1000 with higher rates in Caucasian populations, especially persons of Ashkenazi Jewish ancestry (Calkins et al. 1995; Satsangi et al. 1994). Although the etiology of IBD remains unclear, epidemiological studies have provided evidence of a substantial genetic contribution to IBD susceptibility. Familial aggregation has been observed in 10% of cases, with monozygotic and dizygotic twin concordance rates of 40-50% and 8% respectively, and disease prevalence 10 times greater in first degree relatives of affected persons than in unrelated individuals from the general population (Orholm 1995; Tysk 1988; Mayberry 1980). Presently, IBD is considered a complex genetic disease resulting from interactions between multiple genetic and environmental factors and likely including genetic heterogeneity.

IBD consists of two main clinical forms, Crohn’s disease (CD) and ulcerative colitis (UC). These IBD phenotypes are distinguished by differences in the inflammatory
response with CD affecting any portion of the GI tract, and UC restricted to the large intestine. Certain cases do display features common to both CD and UC and are thus categorized as indeterminate colitis (IC). The CD and UC phenotypes may also coexist within a single family (MIX) further suggesting a shared genetic cause. As there exist both differences and similarities between CD and UC, the genetic basis of IBD is likely to include genes specific to either CD or UC as well as genes common to both CD and UC. This premise is well supported by results of linkage studies which have identified two susceptibility loci specific to CD and several other loci linked to both the CD and UC phenotypes.

Hugot et al. (1996) were first to locate IBD1, a locus on chromosome 16 conferring susceptibility to the CD phenotype. Linkage between IBD1 and CD was subsequently confirmed by Ohmen et al. (1996), Cho et al. (1998) and Hampe et al. (1999a). A second locus, IBD2 on chromosome 12, linked to both CD and UC was identified by Satsangi et al. (1996) and replicated by Hampe et al. (1999a), Duerr et al. (1998) and Ma et al. (1999). A variety of studies have reported suggestive evidence for linkage to IBD at loci on chromosomes 1p, 3p, 3q, 4, 5q, 7, 10 and 22 (Satsangi et al. 1996; Cho et al, 1998; Hampe et al. 1999a 1999b;). Significant evidence for IBD linkage was seen on chromosome 6 (Hampe et al 1999a, 1999b.;), chromosome 14 (Duerr et al. 2000) and chromosome 19 (Riou et al. 2000). Rioux et al. (2000) also report significant linkage between a chromosome region on 5q and CD cases having an early age of disease onset.

The analyses of Rioux et al. (2000) were performed using Canadian IBD (CD, UC and MIX) sibpair families recruited from the Toronto area. Linkage was tested via multipoint sibpair LOD scores using a high density marker map with an average inter-marker distance of 3 cM between markers D5S1470 and D5S1471. Highest evidence for linkage was observed on chromosome 5q in the vicinity of marker D5S2497. Results
of subgroup linkage analyses performed at this locus, were non-significant for UC (LOD=0.1, n=24 families), suggestive for CD (LOD=3.0, n=122) and significant for CD16 (LOD=3.9, n=51) family groups (Rioux et al. 2000). The CD16 subgroup included CD families with at least one affected sibling diagnosed by age ≤ 16. To test whether genetic heterogeneity exists between various IBD family subgroups at this locus, the methods of Chapter 3 were applied to a dataset including a subset of the affected siblings analyzed by Rioux et al. (2000).

## 5.1 The Data

The data analyzed for heterogeneity included 150 IBD (122 CD, 51 CD16 and 28 UC) families recruited as part of the genome-wide screen conducted by Rioux et al. (2000). Diagnoses and age at onset of disease from affected individuals were used to classify families as CD, CD16 or UC following the same criteria employed by Rioux et al. (2000). All families analyzed included at least two individuals affected with either CD or UC and possibly other relatives with IC. The CD> 16 subgroup refers to those CD families not included within the CD16 group (CD = CD16 + CD> 16). In contrast to Rioux et al. (2000) who analyzed affected siblings only, other types of affected relatives, if available, were included in the present analyses. A detailed breakdown of the family structures present is provided in Table 5.2.

Genotyping data were available for 17 markers on chromosome 5 spanning 165 cM with an average inter-marker distance of 10.34 cM (Table 5.1). The marker heterozygosity ranged between 0.62 and 0.84.
Table 5.1: Marker map for chromosome 5.

<table>
<thead>
<tr>
<th>Marker</th>
<th>Distance to next marker (cM)</th>
<th>Distance from first marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>D5S1492</td>
<td>9.6</td>
<td>0</td>
</tr>
<tr>
<td>D5S807</td>
<td>3.9</td>
<td>9.6</td>
</tr>
<tr>
<td>D5S817</td>
<td>3.4</td>
<td>13.5</td>
</tr>
<tr>
<td>D5S1473</td>
<td>9.1</td>
<td>16.9</td>
</tr>
<tr>
<td>D5S1470</td>
<td>13.6</td>
<td>26</td>
</tr>
<tr>
<td>D5S2494</td>
<td>10.3</td>
<td>39.6</td>
</tr>
<tr>
<td>GATA67D03</td>
<td>16.0</td>
<td>49.9</td>
</tr>
<tr>
<td>D5S1501</td>
<td>9.6</td>
<td>65.9</td>
</tr>
<tr>
<td>D5S1719</td>
<td>20.0</td>
<td>75.5</td>
</tr>
<tr>
<td>D5S1453</td>
<td>15.1</td>
<td>95.5</td>
</tr>
<tr>
<td>D5S1505</td>
<td>3.8</td>
<td>110.6</td>
</tr>
<tr>
<td>GATA68A03</td>
<td>5.7</td>
<td>114.4</td>
</tr>
<tr>
<td>D5S816</td>
<td>8.2</td>
<td>120.1</td>
</tr>
<tr>
<td>D5S1480</td>
<td>12.3</td>
<td>128.3</td>
</tr>
<tr>
<td>D5S820</td>
<td>12.3</td>
<td>140.6</td>
</tr>
<tr>
<td>D5S1471</td>
<td>2.7</td>
<td>152.9</td>
</tr>
<tr>
<td>D5S1456</td>
<td>–</td>
<td>155.6</td>
</tr>
</tbody>
</table>
Table 5.2: Breakdown of 150 IBD families analyzed according to the type of affected relatives.

<table>
<thead>
<tr>
<th>Relationship of Other Affected Relatives to the Affected Sibling(s)</th>
<th>Number of Families with</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 or 2 or 3 Affected Siblings</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>-</td>
<td>122</td>
</tr>
<tr>
<td>1 child</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 aunt/uncle</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1 cousin</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1 parent &amp; 1 grandparent</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>1 parent &amp; 1 cousin</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>1 child &amp; 1 nephew/niece</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>1 child &amp; 1 cousin</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>1 great aunt/uncle &amp; 1 great niece</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

5.2 Methods of Analyses

The program PEDMANAGER (Whitehead Institute, 1995) was used to check the genotyping data for Mendelian inconsistencies and to compute allele frequencies among founder individuals. Population allele frequencies, required for *ibd* allele-sharing analyses with incomplete inheritance information, were estimated using the founder allele frequencies computed with PEDMANAGER. Multipoint linkage analyses were conducted at increment steps of 1 cM using the scoring function $S_{all}$ of GENEHUNTER-PLUS (Kruglyak et al. 1996; Kong and Cox 1997) to obtain family-specific NPL scores. Summary NPL statistics were calculated across the chromosome for each of the five family subgroups IBD (CD+UC), CD, CD16, CD> 16 and UC. At the position where the highest summary NPL statistic was observed within the CD group, the distribution of family-specific NPL scores was graphically displayed.
using histograms and boxplots for the five family subgroups.

The linear risk likelihood model of Kong and Cox (1997) was separately fit to each family subgroup to obtain estimates for the linkage parameter $\delta$, and statistic $Z_{tr}$. Heterogeneity analyses, using the methods of Chapter 3, were then conducted at the location where the summary NPL score was highest in CD families. The following comparisons were made: CD vs. UC, CD16 vs. CD $> 16$ and CD16 vs. UC+CD $> 16$. A positive covariate value $X = 1$ was assigned to the CD and CD16 subgroups for which greater linkage evidence was expected. The significance of permutation tests $EST_{perm}$, $T_{UVperm}$ and $T_{EVperm}$ was evaluated using 1000 permutation samples.

### 5.3 Results

The information content of genotyped markers ranged between 0.5 and 0.6 and was similar within all family subgroups analyzed (Figure 5.4). Consistent with the results of Rioux et al. (2000), the greatest evidence for linkage was observed near position cM=130 (Figure 5.1) for both CD and CD16 subgroups. At this locus the summary NPL score increased from NPL=1.86 (p-value=0.0315) for IBD to NPL=2.17 (p-value=0.0150) for CD and NPL=2.82 (p-value=0.0024) for CD16 families. Although the summary NPL score increased, these data do not provide sufficient evidence to declare significant linkage (p-value $< 10^{-5}$ ) for CD16 families according to the criteria of Kruglyak et al. (1995). Figures 5.2 and 5.3 show histograms and boxplots of family-specific NPL scores for IBD, CD, CD16 and UC family subgroups.

The results of linkage and heterogeneity analyses, conducted at cM=130 are given in Table 5.3 (a) and (b). For all family subgroups, except UC, the linkage parameter $\delta$ of the linear risk Kong and Cox (1997) likelihood model maximized within the imposed constraints. Given that the summary NPL score of the UC subgroup is...
Table 5.3: Results of linkage (a) and heterogeneity (b) analyses at position cM=130 on chromosome 5q between IBD family subgroups. \( N_1 \) and \( N_2 \) are the sample sizes of the first and second group listed \( (N = N_1 + N_2) \). \( NPL \) and \( \hat{\delta} \) used to compute \( EST \) and \( EST_{perm} \) correspond to analysis of combined subgroups. \( NPL_1, \hat{\delta}_1 \) and \( NPL_2, \hat{\delta}_2 \) correspond to the first and second family group listed. The estimates \( \hat{\delta} \) and \( \hat{\beta} \) were used to calculate the \( LRT \) statistics. 1000 permutation samples were obtained to compute the significance of \( EST_{perm}, T_{UV_{perm}} \) and \( T_{EV_{perm}} \) tests.

(a)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>( N )</th>
<th>( N_1 )</th>
<th>( N_2 )</th>
<th>( NPL )</th>
<th>( \hat{\delta} )</th>
<th>( NPL_1 )</th>
<th>( \hat{\delta}_1 )</th>
<th>( NPL_2 )</th>
<th>( \hat{\delta}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD vs. UC</td>
<td>150</td>
<td>122</td>
<td>28</td>
<td>1.862</td>
<td>0.204</td>
<td>2.173</td>
<td>0.265</td>
<td>-0.225</td>
<td>0</td>
</tr>
<tr>
<td>CD16 vs. CD&gt;16</td>
<td>122</td>
<td>51</td>
<td>71</td>
<td>2.173</td>
<td>0.265</td>
<td>2.821</td>
<td>0.476</td>
<td>0.457</td>
<td>0.076</td>
</tr>
<tr>
<td>CD16 vs CD&gt;16+UC</td>
<td>150</td>
<td>51</td>
<td>99</td>
<td>1.862</td>
<td>0.204</td>
<td>2.821</td>
<td>0.476</td>
<td>0.267</td>
<td>0.036</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>( \hat{\delta} )</th>
<th>( \hat{\beta} )</th>
<th>( LRT )</th>
<th>( EST )</th>
<th>( EST_{perm} )</th>
<th>( T_{UV} )</th>
<th>( T_{UV_{perm}} )</th>
<th>( T_{EV} )</th>
<th>( T_{EV_{perm}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD vs. UC</td>
<td>0.133</td>
<td>0.133</td>
<td>0.179</td>
<td>0.188</td>
<td>0.215</td>
<td>0.210</td>
<td>0.213</td>
<td>0.183</td>
<td>0.190</td>
</tr>
<tr>
<td>CD16 vs. CD&gt;16</td>
<td>0.276</td>
<td>0.200</td>
<td>0.036</td>
<td>0.031</td>
<td>0.038</td>
<td>0.025</td>
<td>0.029</td>
<td>0.027</td>
<td>0.030</td>
</tr>
<tr>
<td>CD16 vs CD&gt;16+UC</td>
<td>0.256</td>
<td>0.220</td>
<td>0.014</td>
<td>0.011</td>
<td>0.007</td>
<td>0.010</td>
<td>0.005</td>
<td>0.012</td>
<td>0.006</td>
</tr>
</tbody>
</table>

negative and indicates no linkage, it is not surprising that the linkage parameter of the UC group was constrained to the lower bound 0. Within each comparison, heterogeneity methods yielded consistent results (Table 5.3 b). Significant heterogeneity was detected between the CD16 and CD> 16 family groups and between CD16 and combined CD> 16+UC subgroups. Interestingly, no significant heterogeneity was detected for the CD vs. UC comparison.
5.4 Discussion

In a fashion similar to that reported by Rioux et al. (2000), the maximum linkage evidence on chromosome 5 increased when the CD and CD16 families were analyzed separately. However, the observed increase was not sufficient to declare significant linkage for CD16 using these data. It is not surprising that the current data provide less evidence for linkage than observed by Rioux et al. (2000) considering the lower density of genetic markers used in these analyses. The average inter-marker distance for the current data was > 10 cM. In contrast Rioux et al. (2000) obtained a significant linkage result for CD16 families using a dense marker map with an additional 34 markers around the peak region and average inter-marker distance of 3 cM. Another important difference between the present analyses and those of Rioux et al. (2000) is that here *ibd* allele sharing was considered for sets of affected relatives. The analyses of Rioux et al. (2000) included only affected sibling pairs. The heterogeneity methods examined in this thesis were specifically designed to utilize *ibd* allele-sharing evidence from sets of affected relatives including the case of affected sibpairs.

The linkage parameter $\delta$ in the Kong and Cox (1997) linear risk model maximized within the imposed bounds for all subgroups, except UC (Table 5.3 a). It is uncertain how closely the test statistics $LRT$ and $EST$ followed their respective asymptotic distributions, but within each comparison, the results of heterogeneity tests were consistent. With the available data, all methods had sufficient power to detect heterogeneity at a 5% testing level between the CD16 and CD $>$ 16 groups and between CD16 and CD $>$ 16 + UC. No heterogeneity was detected between CD and UC families although the difference between summary NPL scores of CD and UC was comparable to the other two comparisons. We suspect that there is insufficient power to detect heterogeneity between CD ($n=122$) and UC ($n=28$) because the UC sample
size is too small.
Figure 5.1: Summary multipoint NPL scores for the IBD (CD+UC, n=150 families), CD (n=122), CD16 (n=51), CD > 16 (n=71) and UC (n=28) family subgroups calculated using the scoring function $S_{df}$ implemented in GENEHUNTER-PLUS.
Figure 5.2: Histogram of family-specific NPL scores at cM=150 for IBD (CD+UC), CD, and UC (n=152), CD16 (n=51), and UC (n=28) family subgroups.
Figure 5.3: Boxplots of family-specific NPL scores at cM=130 for IBD (CD+UC, n=150 families), CD (n=122), CD16 (n=51), CD > 16 (n=71) and UC (n=28) family subgroups.
Figure 5.4: Information content for IBD (CD+UC, n=150 families), CD (n=122), CD16 (n=51), CD> 16 (n=71) and UC (n=28) family subgroups.
Chapter 6

Conclusions and Discussion

The methods \( LRT, EST, T_{UV}, T_{UVperm}, T_{EV} \) and \( T_{EVperm} \) were developed to assess the presence of genetic heterogeneity within the context of linkage studies. The corresponding test statistics were computed using the same family-specific binary covariate \( X \) and the same NPL scores that for each family measure excess \( ibd \) allele sharing among affected relatives. In essence, these methods consider differences between the observed allele-sharing distribution of the two covariate-defined family groups. An important issue is how well the covariate selected identifies heterogeneity in the allele-sharing distribution of family subgroups. Also important are factors that may change the sample distribution of the covariate \( X \) or NPL scores upon which the heterogeneity tests depend. These include mode of disease inheritance, kinship among affected relatives, extent of linkage and size of family subgroups.

For each family, the NPL score was computed as the expected value of a normalized score of the function \( S_{all} \) (2.2). In the absence of linkage, NPL scores, calculated using complete inheritance information, have mean 0 and variance 1. When the data contains pedigrees with linkage, the distribution of NPL scores depends on the extent of linkage present. We generated data sets containing tight, moderate and no linkage under the null hypothesis of no heterogeneity. Although at each linkage
level, the same distribution of NPL scores was present in both pedigree subgroups, the distribution of test statistics \( LRT \), \( EST \), \( T_{UV} \) and \( T_{EV} \) varied between linkage levels. Therefore, the significance of these test statistics cannot be assessed using the expected distribution specified by asymptotic likelihood theory. Our suggestion is to assess significance using distribution-free approaches, such as the permutation methods applied to \( EST_{perm} \), \( T_{UVperm} \) and \( T_{EVperm} \). Note that even if permutation methods are used to assess significance, maximum likelihood estimates must be obtained to compute test statistics derived from the likelihood model (3.5).

Regardless of how significance is determined, the extent of linkage within family subgroups is one factor that is likely to affect the power of methods to detect heterogeneity. Our simulation study considered only the ideal case in which linkage was either completely absent or completely present. Also, simulations restricted the family covariate \( X \) such that linked and unlinked family subgroups were precisely identified. We expect the power of heterogeneity tests to decrease in the presence of covariate misclassification and/or if family subgroups contain intermediate linkage levels.

Mode of disease inheritance and kinship of affected relatives are two other factors that may influence the distribution of NPL scores and subsequently affect the power of heterogeneity tests. McPeek (1999) showed that the scoring function \( S_{alt} \) is optimal for an additive model of disease inheritance and that the value of an optimal scoring function does not depend on whether sharing is between more closely or more distantly related affected individuals. Our simulation study was designed to examine the power of heterogeneity tests under dominant and recessive inheritance models for which the scoring function \( S_{alt} \) is not optimal. In general, the power of heterogeneity tests was lower for the recessive model than for the dominant. This is consistent with simulation results from both Kruglyak et al. (1996) and McPeak (1999) which showed that the
scoring function \( S_{all} \) has greater power to detect linkage under dominant rather than recessive inheritance. A question not addressed by our simulation study, is the type of kinship relationships, specific to each inheritance model, that will render heterogeneity methods more powerful.

Power comparisons for \( LRT^* \), \( T_{UVperm} \) and \( T_{EVperm} \) test statistics applied to simulated data sets showed reduced power when the overall sample size was decreased. A reduction in power was also observed as the sizes of family subgroups became less balanced. In the presence of heterogeneity, data sets with unequal subgroups were simulated such that the larger subgroup contained linked families. When the reverse is true, we do not expect any changes in the performance of \( T_{UV} \), \( T_{UVperm} \), \( T_{EV} \) and \( T_{EVperm} \), but the behaviour of the likelihood-based methods is more difficult to predict. The importance of the covariate distribution is clearly demonstrated by the IBD example in Chapter 5, for which methods failed to detect significant heterogeneity when the small UC sample size was analyzed. We emphasize once again that the covariate \( X \) should be selected carefully because the power of methods to detect heterogeneity will depend on how well \( X \) identifies heterogeneity among the allele-sharing distribution of sampled families.

Other methods may be considered to test for heterogeneity using family-specific covariate classes and NPL linkage scores. One possibility is to use the covariate \( X \) as the response variable in a logistic regression model with NPL scores as predictor values. An advantage of this approach is that no assumptions about the distribution of NPL scores are required. The heterogeneity tests we have examined may be generalized to allow for more than two covariate classes. Alternatively, regression methods for ordinal responses \( X \) may be used to test for heterogeneity among multiple covariate classes. However, the issues discussed above will persist for any approach that is based on family-specific covariate classes and NPL linkage scores.
Testing level is also a relevant concern of any method that tests for heterogeneity in observed linkage scores. In linkage analysis, a stringent testing level \((p\text{-value} < 10^{-5})\) (Lander and Kruglyak, 1995) is used that indicates the probability of obtaining a significant linkage result, by chance alone, anywhere within the genome. In this thesis, we have evaluated the significance of heterogeneity tests using a nominal 5% critical level, without consideration of multiple testing. Further investigation is needed to determine what is an appropriate testing level that, given a data set collected for a genome-wide linkage study, considers when, how and the number of heterogeneity analyses performed.

Amongst the methods examined in this thesis, we recommend the use of \(T_{EV\text{perm}}\) to test for heterogeneity between the allele-sharing distributions of family subgroups. Under all scenarios examined, \(T_{EV\text{perm}}\) was found to have a nominal 5% error rate and power, \(> 90\%\), greater than or comparable to that of the other methods evaluated. We do caution that the power of \(T_{EV\text{perm}}\) depends on the size of family subgroups defined. In future research we aim to further investigate \(T_{EV\text{perm}}\) considering a broad range of linkage levels and subgroup sample sizes. Also of interest is the comparison of \(T_{EV\text{perm}}\) to other genetic heterogeneity tests such as the methods developed by Greenwood and Bull (1999) for sibpairs and those of Olson (1999) that consider allele sharing between pairs of any type of affected relatives.
Appendix A

Derivation of Certain Results

A.1 Calculation of Maximum Likelihood Estimates for Parameters $\delta$ and $\beta$

The natural logarithm of the likelihood function (3.5) with weights $\gamma_i = 1$ is written as follows considering the two pedigree subgroups $g_1$ and $g_2$,

$$l(t, \delta, \beta) = C + \sum_{i=1}^{N} \ln[1 + \delta Z_i(t) + \beta X_i Z_i(t)]$$

$$= C + \sum_{g_1} \ln[1 + \delta_1 Z_i(t)] + \sum_{g_2} \ln[1 + \delta_2 Z_i(t)]$$

$$= l(t, \delta_1, \delta_2)$$

(A.1)

where $\delta_1 = \delta + \beta$ and $\delta_2 = \delta - \beta$. Maximum likelihood estimates $\hat{\delta}$ and $\hat{\beta}$ do not need to be iteratively computed, but can be obtained as

$$\hat{\delta} = \frac{\delta_1 + \delta_2}{2}$$

(A.2)

and

$$\hat{\beta} = \frac{\delta_1 - \delta_2}{2}$$

(A.3)

using the subgroup maximum likelihood estimates $\hat{\delta}_1$ and $\hat{\delta}_2$. 
To prove that the above $\bar{\delta}$ and $\bar{\beta}$ estimates do maximize the likelihood function in (3.5), we show that the score vector $U(t, \delta, \beta)$ evaluated using these estimates is 0.

The subgroup maximum likelihood estimates $\hat{\delta}_1$ and $\hat{\delta}_2$ satisfy

$$\left. \frac{\partial U(t, \delta_1, \delta_2)}{\partial \delta_1} \right|_{\delta_1 = \hat{\delta}_1} = \sum_{g_1} \frac{Z_i(t)}{1 + \hat{\delta}_1 Z_i(t)}$$

and

$$\left. \frac{\partial U(t, \delta_1, \delta_2)}{\partial \delta_2} \right|_{\delta_2 = \hat{\delta}_2} = \sum_{g_2} \frac{Z_i(t)}{1 + \hat{\delta}_2 Z_i(t)}$$

The components of the score vector, $U_1(t, \delta, \beta)$ and $U_2(t, \delta, \beta)$, evaluated at $\bar{\delta}$ and $\bar{\beta}$ are derived as:

$$U_1(t, \bar{\delta}, \bar{\beta}) = \left. \frac{\partial U(t, \delta, \beta)}{\partial \delta} \right|_{\delta = \bar{\delta}, \beta = \bar{\beta}}$$

$$= \sum_{i=1}^N \frac{Z_i(t)}{1 + \bar{\delta} Z_i(t) + \bar{\beta} X_i Z_i(t)}$$

$$= \sum_{g_1} \frac{Z_i(t)}{1 + \hat{\delta}_1 Z_i(t)} + \sum_{g_2} \frac{Z_i(t)}{1 + \hat{\delta}_2 Z_i(t)}$$

$$= 0$$

$$U_2(t, \bar{\delta}, \bar{\beta}) = \left. \frac{\partial U(t, \delta, \beta)}{\partial \beta} \right|_{\delta = \bar{\delta}, \beta = \bar{\beta}}$$

$$= \sum_{i=1}^N \frac{X_i Z_i(t)}{1 + \bar{\delta} Z_i(t) + \bar{\beta} X_i Z_i(t)}$$

$$= \sum_{g_1} \frac{Z_i(t)}{1 + \hat{\delta}_1 Z_i(t)} - \sum_{g_2} \frac{Z_i(t)}{1 + \hat{\delta}_2 Z_i(t)}$$

$$= 0$$

93
Since the components of $U(t, \tilde{\delta}, \tilde{\beta})$ are both equal to 0, we have proved that the $\tilde{\delta}$ and $\tilde{\beta}$ defined in (A.2) and (A.3) are indeed the maximum likelihood estimates.

### A.2 Constraints for the Maximum Likelihood Estimates $\tilde{\delta}, \tilde{\beta}$

The maximum likelihood estimates $\tilde{\delta}$ (A.2) and $\tilde{\beta}$ (A.3) are derived using the subgroup maximum likelihood estimates $\hat{\delta}_1$ and $\hat{\delta}_2$ which are subject to the following constraints.

\[ 0 \leq \hat{\delta}_1 \leq b_1 \]  
\[ (A.8) \]

\[ 0 \leq \hat{\delta}_2 \leq b_2 \]  
\[ (A.9) \]

For pedigree groups of size $N_1$ and $N_2$, the upper bounds are obtained as $b_1 = \min(b_i)$, $i = 1 \ldots N_1$ and $b_2 = \min(b_i)$, $i = 1 \ldots N_2$. The function $b_i(t)$

\[ b_i(t) = \frac{\sigma_i}{(\mu_i - a_i(t))} \]  
\[ (A.10) \]

is defined specific to locus $t$ and pedigree $i$ in terms of the mean $\mu_i$, standard deviation $\sigma_i$ and smallest possible value $a_i(t)$ of the scoring function $S_i(t)$. Under the null hypothesis of no heterogeneity, the maximum likelihood estimate, $0 \leq \hat{\delta} \leq b$, has upper bound $b = \min(b_1, b_2)$. The maximum likelihood estimate $\tilde{\delta}$ is therefore constrained to

\[ 0 \leq \tilde{\delta} \leq \min(b_1, b_2) = b \]  
\[ (A.11) \]
To ensure a positive phenotype risk ratio, the following inequality must hold

\[
R_i(t, \delta, \beta) \geq 0 \\
\left[ 1 + \left( \frac{S_i(t) - \mu_i}{\sigma_i} \right) \delta + \left( \frac{S_i(t) - \mu_i}{\sigma_i} \right) X_i\beta \right] \geq 0 \\
(a_i(t) - \mu_i) (\delta + X_i\beta) \geq -\sigma_i \\
(\delta + X_i\beta) \leq \frac{-\sigma_i}{a_i(t) - \mu_i} \\
X_i\beta \leq \left( \frac{\sigma_i}{\mu_i - a_i(t)} \right) - \delta \\
X_i\beta \leq b_i(t) - \delta,
\]

and depending on whether the value of \( X_i \) is 1 or -1, \( \beta \) must satisfy both of the following inequalities.

\[
\begin{cases} 
\beta \leq b - \delta & \text{if } X_i = 1 \\
\beta \geq \delta - b & \text{if } X_i = -1
\end{cases}
\] (A.12)

Therefore \( \bar{\beta} \) is constrained to

\[
\bar{\delta} - b \leq \bar{\beta} \leq b - \bar{\delta}.
\] (A.13)

### A.3 Results When the Parameter \( \delta \) Maximize Within the Interval \([0, b]\)

The expressions derived below are in reference to \( l(t, \hat{\delta}, \beta_0) \) of (A.1) and are required in the development of the score statistic \( EST_1(t, \hat{\delta}, \beta_0) \) (3.27). The expected value is taken with respect to the null hypothesis of no heterogeneity when \( \beta = \beta_0 \) and \( \hat{\delta} \) is the maximum likelihood estimate evaluated when \( \beta_0 = 0 \). As in Chapter 3,

\[
u_i(t) = \frac{Z_i(t)}{1 + \hat{\delta} Z_i(t)}
\] (A.14)
\[ U_1(t, \delta, \beta_0) = \frac{\partial U(t, \delta, \beta)}{\partial \delta} \bigg|_{\delta=\delta}^{\beta=\beta_0} = \sum_{i=1}^{N} \frac{Z_i(t)}{1 + \delta Z_i(t)} = \sum_{i=1}^{N} u_i(t) \]  
(A.15)

\[ U_2(t, \delta, \beta_0) = \frac{\partial l(t, \delta, \beta)}{\partial \beta} \bigg|_{\delta=\delta}^{\beta=\beta_0} = \sum_{i=1}^{N} \frac{X_i Z_i(t)}{1 + \delta Z_i(t)} = \sum_{i=1}^{N} X_i u_i(t) \]  
(A.16)

Let

\[ x = -E_0 \left[ \frac{\partial^2 l(t, \delta, \beta)}{\partial \delta^2} \right]_{\delta=\delta}^{\beta=\beta_0} = -E_0 \left[ \frac{\partial^2 l(t, \delta, \beta)}{\partial \beta^2} \right]_{\delta=\delta}^{\beta=\beta_0} = E_0 \left[ \sum_{i=1}^{N} \left( \frac{Z_i(t)}{1 + \delta Z_i(t)} \right)^2 \right] = E_0 \left[ \sum_{i=1}^{N} u_i^2(t) \right] \]  
(A.17)

\[ y = -E_0 \left[ \frac{\partial^2 l(t, \delta, \beta)}{\partial \delta \partial \beta} \right]_{\delta=\delta}^{\beta=\beta_0} \]
The score vector, $U(t, \delta, \beta_0)$, corresponding to the null hypothesis of no heterogeneity has components defined in (A.15) and (A.16).

$$U(t, \delta, \beta_0) = \begin{pmatrix} U_1(t, \delta, \beta) \\ U_2(t, \delta, \beta) \end{pmatrix}$$  \hspace{1cm} (A.19)

As outlined in Chapter 3, when the value of $\delta$ that maximizes (2.31) lies within the constrained parameter space $[0, b]$, then the first component of the score vector equals 0. Therefore,

$$E_0[U_1(t, \delta, \beta)] = E_0[0]$$

$$E_0 \left[ \sum_{i=1}^{N} \frac{Z_i(t)}{1 + \delta Z_i(t)} \right] = 0$$

Under the null hypothesis of no heterogeneity, the pedigree-specific NPL scores $Z_i(t)$ are assumed independent and identically distributed such that

$$E_0 \left[ \frac{Z_i(t)}{1 + \delta Z_i(t)} \right] = 0$$  \hspace{1cm} (A.20)

It then follows that

97
\[ E_0[U_2(t, \delta, \beta)] = E_0 \left[ \sum_{i=1}^{N} \frac{X_i Z_i(t)}{1 + \delta Z_i(t)} \right] \]

\[ = \sum_{i=1}^{N} X_i E_0 \left[ \frac{Z_i(t)}{1 + \delta Z_i(t)} \right] \]

\[ = 0 \]

and, therefore, the null expected value of the score vector is 0.

\[ E_0[U(t, \delta, \beta_0)] = 0 \quad (A.21) \]

When (A.21) holds true, the variance-covariance matrix \( \Sigma_0(t, \delta, \beta_0) \) of the score vector evaluated under the null hypothesis of no heterogeneity is

\[ \Sigma_0(t, \delta, \beta_0) = \text{Cov}_0 \left[ U(t, \delta, \beta_0), U(t, \delta, \beta_0) \right] \]

\[ = \begin{pmatrix}
\text{Var}_0[U_1(t, \delta, \beta_0)] & \text{Cov}_0[U_1(t, \delta, \beta_0), U_2(t, \delta, \beta_0)] \\
\text{Cov}_0[U_2(t, \delta, \beta_0), U_1(t, \delta, \beta_0)] & \text{Var}_0[U_2(t, \delta, \beta_0)]
\end{pmatrix} \]

which is evaluated by

\[ \Sigma_0(t, \delta, \beta_0) = -E_0 \left( \begin{pmatrix}
\frac{\partial^2 l(t, \delta, \beta)}{\partial \delta^2} & \frac{\partial^2 l(t, \delta, \beta)}{\partial \delta \partial \beta} \\
\frac{\partial^2 l(t, \delta, \beta)}{\partial \beta \partial \delta} & \frac{\partial^2 l(t, \delta, \beta)}{\partial \beta^2}
\end{pmatrix} \right) \]

\[ \delta=\delta_0 \]

\[ \beta=\beta_0 \]

\[ = \begin{pmatrix} x & y \\ y & x \end{pmatrix} \]

for \( x \) and \( y \) as defined in (A.17) and (A.18).
### A.4 Generation of Sibships ≥ 2

The size of shipships was modelled by Suarez and Van Eerdewegh (1984) as a random variable $X$ from a geometric distribution ($p = 0.4551$) with probability density function (pdf)

$$
P(X = x) = p(1 - p)^{x-1} \quad x = 1, 2, 3, \ldots \quad (A.23)
$$

To generate sibships containing 2 or more individuals, $X$ must follow a geometric distribution left truncated at 2 with pdf

$$
P(X = x) = \frac{p(1 - p)^{x-1}}{\sum_{i=2}^{\infty} p(1 - p)^{i-1}} \quad x = 2, 3, \ldots \quad (A.24)
$$

When both numerator and denominator are multiplied by a factor $(1 - p)^{-1}$, then

$$
P(X = x) = \frac{p(1 - p)^{x-1}(1 - p)^{-1}}{\sum_{i=2}^{\infty} p(1 - p)^{i-1}(1 - p)^{-1}} \quad x = 2, 3, \ldots
$$

$$
= \frac{p(1 - p)^{x-2}}{\sum_{i=1}^{\infty} p(1 - p)^{i-1}}
$$

$$
= p(1 - p)^{x-2} \quad (A.25)
$$

since $\sum_{i=1}^{\infty} p(1 - p)^{i-1} = 1$. Let $Y = X - 2$, then by (A.25),

$$
P(Y = y) = p(1 - p)^y \quad y = 0, 1, 2, \ldots
$$

is the pdf of a random variable $Y$ from a negative binomial distribution with parameters $r = 1$ and $p$. The general formula for the pdf of random variable $Y$ with negative binomial distribution $(r, p)$ is:

$$
P(Y = y) = \binom{r + y - 1}{y} p^r (1 - p)^y \quad y = 0, 1, 2, \ldots \quad (A.26)
$$
Therefore, using the relation $X = Y + 2$, sibships of size $\geq 2$ are obtained by adding 2 to a random variate generated from a negative binomial distribution with parameters $r = 1, p = 0.4551$. 
Appendix B

Results of Simulations

The results of the simulation study described in Chapter 4 are summarized in Tables B.1-B.4 within this appendix. Test statistics examined include $LRT$, $EST$, $EST_1$, $T_{EV}$ and $T_{UV}$ for which significance was computed using the expected asymptotic distribution, permutation methods $EST_{perm}$, $EST_{1perm}$, $T_{EVperm}$ and $T_{UVperm}$ and $LRT^*$, $EST^*$, $EST_1^*$, $T_{EV}$ and $T_{UV}$ tests with significance based on empirical 5% critical values. Results consist of type I error estimates, power estimates and corresponding 95% confidence levels. Replicates were generated in the presence and absence of locus heterogeneity using dominant and recessive inheritance models. Pedigree samples included either a total of 150 families (SAMP) or a minimum of 75 linked pedigrees (LINK). Families were classified according to a binary covariate $X = 1, -1$ with a proportion $\alpha$ (0.8, 0.6 and 0.5) of families with covariate value 1.
Table B.1: Type I error estimates and 95% confidence intervals calculated for SAMP and LINK α-samples under dominant inheritance. Reps is the number of replicates performed under the null hypothesis of no heterogeneity.

<table>
<thead>
<tr>
<th>(α)</th>
<th>Test</th>
<th>Reps</th>
<th>SAMP</th>
<th></th>
<th>Repe</th>
<th>LINK</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Error</td>
<td>95% CI</td>
<td></td>
<td>Error</td>
<td>95% CI</td>
</tr>
<tr>
<td>0.8</td>
<td>LRT</td>
<td>1000</td>
<td>0.003</td>
<td>0.001</td>
<td>0.010</td>
<td>1000</td>
<td>0.001</td>
</tr>
<tr>
<td>0.8</td>
<td>EST1</td>
<td>267</td>
<td>0.930</td>
<td>0.880</td>
<td>0.950</td>
<td>171</td>
<td>0.820</td>
</tr>
<tr>
<td>0.8</td>
<td>EST1perm</td>
<td>267</td>
<td>0.050</td>
<td>0.030</td>
<td>0.080</td>
<td>171</td>
<td>0.050</td>
</tr>
<tr>
<td>0.8</td>
<td>EST</td>
<td>1000</td>
<td>0.980</td>
<td>0.970</td>
<td>0.990</td>
<td>1000</td>
<td>0.970</td>
</tr>
<tr>
<td>0.8</td>
<td>ESTperm</td>
<td>1000</td>
<td>0.050</td>
<td>0.040</td>
<td>0.070</td>
<td>1000</td>
<td>0.050</td>
</tr>
<tr>
<td>0.8</td>
<td>Tev</td>
<td>1000</td>
<td>0.012</td>
<td>0.007</td>
<td>0.021</td>
<td>1000</td>
<td>0.019</td>
</tr>
<tr>
<td>0.8</td>
<td>Tevperm</td>
<td>1000</td>
<td>0.036</td>
<td>0.026</td>
<td>0.050</td>
<td>1000</td>
<td>0.037</td>
</tr>
<tr>
<td>0.8</td>
<td>Tuv</td>
<td>1000</td>
<td>0.014</td>
<td>0.008</td>
<td>0.024</td>
<td>1000</td>
<td>0.022</td>
</tr>
<tr>
<td>0.8</td>
<td>Tuvperm</td>
<td>1000</td>
<td>0.036</td>
<td>0.026</td>
<td>0.050</td>
<td>1000</td>
<td>0.033</td>
</tr>
<tr>
<td>0.6</td>
<td>LRT</td>
<td>1000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.006</td>
<td>1000</td>
<td>0.001</td>
</tr>
<tr>
<td>0.6</td>
<td>EST1</td>
<td>267</td>
<td>0.930</td>
<td>0.890</td>
<td>0.960</td>
<td>230</td>
<td>0.930</td>
</tr>
<tr>
<td>0.6</td>
<td>EST1perm</td>
<td>267</td>
<td>0.030</td>
<td>0.010</td>
<td>0.060</td>
<td>230</td>
<td>0.040</td>
</tr>
<tr>
<td>0.6</td>
<td>EST</td>
<td>1000</td>
<td>0.980</td>
<td>0.970</td>
<td>0.990</td>
<td>1000</td>
<td>0.980</td>
</tr>
<tr>
<td>0.6</td>
<td>ESTperm</td>
<td>1000</td>
<td>0.050</td>
<td>0.040</td>
<td>0.070</td>
<td>1000</td>
<td>0.050</td>
</tr>
<tr>
<td>0.6</td>
<td>Tev</td>
<td>1000</td>
<td>0.015</td>
<td>0.009</td>
<td>0.025</td>
<td>1000</td>
<td>0.022</td>
</tr>
<tr>
<td>0.6</td>
<td>Tevperm</td>
<td>1000</td>
<td>0.039</td>
<td>0.028</td>
<td>0.053</td>
<td>1000</td>
<td>0.046</td>
</tr>
<tr>
<td>0.6</td>
<td>Tuv</td>
<td>1000</td>
<td>0.016</td>
<td>0.009</td>
<td>0.026</td>
<td>1000</td>
<td>0.025</td>
</tr>
<tr>
<td>0.6</td>
<td>Tuvperm</td>
<td>1000</td>
<td>0.037</td>
<td>0.027</td>
<td>0.051</td>
<td>1000</td>
<td>0.044</td>
</tr>
<tr>
<td>0.5</td>
<td>LRT</td>
<td>1000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.006</td>
<td>1000</td>
<td>0.001</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1</td>
<td>267</td>
<td>0.920</td>
<td>0.880</td>
<td>0.950</td>
<td>267</td>
<td>0.920</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1perm</td>
<td>267</td>
<td>0.040</td>
<td>0.020</td>
<td>0.070</td>
<td>267</td>
<td>0.060</td>
</tr>
<tr>
<td>0.5</td>
<td>EST</td>
<td>1000</td>
<td>0.980</td>
<td>0.970</td>
<td>0.990</td>
<td>1000</td>
<td>0.980</td>
</tr>
<tr>
<td>0.5</td>
<td>ESTperm</td>
<td>1000</td>
<td>0.060</td>
<td>0.040</td>
<td>0.070</td>
<td>1000</td>
<td>0.060</td>
</tr>
<tr>
<td>0.5</td>
<td>Tev</td>
<td>1000</td>
<td>0.027</td>
<td>0.018</td>
<td>0.040</td>
<td>1000</td>
<td>0.027</td>
</tr>
<tr>
<td>0.5</td>
<td>Tevperm</td>
<td>1000</td>
<td>0.051</td>
<td>0.039</td>
<td>0.067</td>
<td>1000</td>
<td>0.047</td>
</tr>
<tr>
<td>0.5</td>
<td>Tuv</td>
<td>1000</td>
<td>0.027</td>
<td>0.018</td>
<td>0.040</td>
<td>1000</td>
<td>0.027</td>
</tr>
<tr>
<td>0.5</td>
<td>Tuvperm</td>
<td>1000</td>
<td>0.051</td>
<td>0.039</td>
<td>0.067</td>
<td>1000</td>
<td>0.047</td>
</tr>
</tbody>
</table>
Table B.2: Power estimates and 95% confidence intervals calculated for SAMP and LINK α-samples under dominant inheritance. Reps is the number of replicates performed in the presence of heterogeneity.

<table>
<thead>
<tr>
<th>(a) Test</th>
<th>Reps</th>
<th>SAMP</th>
<th>95% CI</th>
<th>Power</th>
<th>Reps</th>
<th>LINK</th>
<th>95% CI</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6 LRT</td>
<td>568</td>
<td>0.984</td>
<td>0.969</td>
<td>0.992</td>
<td>568</td>
<td>0.914</td>
<td>0.887</td>
<td>0.935</td>
</tr>
<tr>
<td>0.6 LRT*</td>
<td>568</td>
<td>0.998</td>
<td>0.989</td>
<td>1.000</td>
<td>568</td>
<td>0.974</td>
<td>0.956</td>
<td>0.985</td>
</tr>
<tr>
<td>0.6 EST1</td>
<td>553</td>
<td>0.980</td>
<td>0.960</td>
<td>0.990</td>
<td>561</td>
<td>0.960</td>
<td>0.940</td>
<td>0.970</td>
</tr>
<tr>
<td>0.6 EST1perm</td>
<td>553</td>
<td>0.900</td>
<td>0.870</td>
<td>0.920</td>
<td>561</td>
<td>0.890</td>
<td>0.860</td>
<td>0.910</td>
</tr>
<tr>
<td>0.6 EST1*</td>
<td>553</td>
<td>0.416</td>
<td>0.375</td>
<td>0.458</td>
<td>561</td>
<td>0.455</td>
<td>0.413</td>
<td>0.497</td>
</tr>
<tr>
<td>0.6 EST</td>
<td>568</td>
<td>0.980</td>
<td>0.960</td>
<td>0.990</td>
<td>568</td>
<td>0.960</td>
<td>0.940</td>
<td>0.970</td>
</tr>
<tr>
<td>0.6 ESTperm</td>
<td>568</td>
<td>0.900</td>
<td>0.870</td>
<td>0.920</td>
<td>568</td>
<td>0.890</td>
<td>0.860</td>
<td>0.910</td>
</tr>
<tr>
<td>0.6 EST*</td>
<td>568</td>
<td>0.011</td>
<td>0.004</td>
<td>0.024</td>
<td>568</td>
<td>0.004</td>
<td>0.001</td>
<td>0.014</td>
</tr>
<tr>
<td>0.6 Tev</td>
<td>568</td>
<td>0.998</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>0.977</td>
<td>0.965</td>
<td>0.985</td>
</tr>
<tr>
<td>0.6 Tevperm</td>
<td>568</td>
<td>0.998</td>
<td>0.998</td>
<td>1.000</td>
<td>568</td>
<td>0.981</td>
<td>0.965</td>
<td>0.990</td>
</tr>
<tr>
<td>0.6 Tev*</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>0.991</td>
<td>0.978</td>
<td>0.997</td>
</tr>
<tr>
<td>0.6 Tev</td>
<td>568</td>
<td>0.995</td>
<td>0.983</td>
<td>0.999</td>
<td>568</td>
<td>0.964</td>
<td>0.933</td>
<td>0.969</td>
</tr>
<tr>
<td>0.6 Tevperm</td>
<td>568</td>
<td>0.996</td>
<td>0.986</td>
<td>0.999</td>
<td>568</td>
<td>0.954</td>
<td>0.933</td>
<td>0.969</td>
</tr>
<tr>
<td>0.6 Tev*</td>
<td>568</td>
<td>0.998</td>
<td>0.989</td>
<td>1.000</td>
<td>568</td>
<td>0.982</td>
<td>0.967</td>
<td>0.991</td>
</tr>
<tr>
<td>0.6 LRT</td>
<td>568</td>
<td>0.998</td>
<td>0.989</td>
<td>1.000</td>
<td>568</td>
<td>0.998</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 LRT*</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 EST1</td>
<td>558</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
<td>566</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 EST1perm</td>
<td>558</td>
<td>0.990</td>
<td>0.980</td>
<td>1.000</td>
<td>556</td>
<td>0.990</td>
<td>0.980</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 EST1*</td>
<td>558</td>
<td>0.477</td>
<td>0.435</td>
<td>0.519</td>
<td>566</td>
<td>0.485</td>
<td>0.453</td>
<td>0.537</td>
</tr>
<tr>
<td>0.6 EST</td>
<td>568</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 ESTperm</td>
<td>568</td>
<td>0.990</td>
<td>0.980</td>
<td>1.000</td>
<td>568</td>
<td>0.990</td>
<td>0.980</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 EST*</td>
<td>568</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>568</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.6 Tev</td>
<td>568</td>
<td>1.000</td>
<td>0.995</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.995</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 Tevperm</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 Tev*</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 Tev</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 Tevperm</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.6 Tev*</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table B.2: continued

<table>
<thead>
<tr>
<th>(a)</th>
<th>Test</th>
<th>Reps</th>
<th>SAMP</th>
<th>Reps</th>
<th>LINK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>LRT</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>LRT*</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1</td>
<td>568</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1perm</td>
<td>568</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1*</td>
<td>568</td>
<td>0.417</td>
<td>0.377</td>
<td>0.459</td>
</tr>
<tr>
<td>0.5</td>
<td>EST</td>
<td>568</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>ESTperm</td>
<td>568</td>
<td>1.000</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>EST*</td>
<td>568</td>
<td>0.000</td>
<td>0.000</td>
<td>0.008</td>
</tr>
<tr>
<td>0.5</td>
<td>Tsv</td>
<td>568</td>
<td>1.000</td>
<td>0.995</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>Tsvperm</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>Tsv*</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>Tsv</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>Tsvperm</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>Tsv*</td>
<td>568</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table B.3: Type I error estimates and 95% confidence intervals calculated for SAMP and LINK α-samples under recessive inheritance. Reps is the number of replicates performed under the null hypothesis of no heterogeneity.

<table>
<thead>
<tr>
<th>(a) Test</th>
<th>Reps</th>
<th>SAMP</th>
<th>Error</th>
<th>95% CI</th>
<th>Reps</th>
<th>Error</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRT</td>
<td>1000</td>
<td></td>
<td>0.026</td>
<td>0.017</td>
<td>0.038</td>
<td>1000</td>
<td>0.031</td>
</tr>
<tr>
<td>EST</td>
<td>838</td>
<td></td>
<td>0.330</td>
<td>0.300</td>
<td>0.370</td>
<td>750</td>
<td>0.300</td>
</tr>
<tr>
<td>ESTperm</td>
<td>838</td>
<td></td>
<td>0.050</td>
<td>0.040</td>
<td>0.070</td>
<td>750</td>
<td>0.040</td>
</tr>
<tr>
<td>EST</td>
<td>1000</td>
<td></td>
<td>0.400</td>
<td>0.370</td>
<td>0.430</td>
<td>1000</td>
<td>0.410</td>
</tr>
<tr>
<td>ESTperm</td>
<td>1000</td>
<td></td>
<td>0.040</td>
<td>0.030</td>
<td>0.060</td>
<td>1000</td>
<td>0.040</td>
</tr>
<tr>
<td>Tev</td>
<td>1000</td>
<td></td>
<td>0.025</td>
<td>0.017</td>
<td>0.037</td>
<td>1000</td>
<td>0.043</td>
</tr>
<tr>
<td>Tevperm</td>
<td>1000</td>
<td></td>
<td>0.058</td>
<td>0.045</td>
<td>0.075</td>
<td>1000</td>
<td>0.064</td>
</tr>
<tr>
<td>Tuv</td>
<td>1000</td>
<td></td>
<td>0.027</td>
<td>0.018</td>
<td>0.040</td>
<td>1000</td>
<td>0.042</td>
</tr>
<tr>
<td>Tuvperm</td>
<td>1000</td>
<td></td>
<td>0.057</td>
<td>0.044</td>
<td>0.074</td>
<td>1000</td>
<td>0.060</td>
</tr>
<tr>
<td>LRT</td>
<td>1000</td>
<td></td>
<td>0.052</td>
<td>0.039</td>
<td>0.068</td>
<td>1000</td>
<td>0.040</td>
</tr>
<tr>
<td>EST</td>
<td>838</td>
<td></td>
<td>0.350</td>
<td>0.320</td>
<td>0.390</td>
<td>810</td>
<td>0.370</td>
</tr>
<tr>
<td>ESTperm</td>
<td>838</td>
<td></td>
<td>0.050</td>
<td>0.040</td>
<td>0.070</td>
<td>810</td>
<td>0.050</td>
</tr>
<tr>
<td>EST</td>
<td>1000</td>
<td></td>
<td>0.410</td>
<td>0.380</td>
<td>0.440</td>
<td>1000</td>
<td>0.440</td>
</tr>
<tr>
<td>ESTperm</td>
<td>1000</td>
<td></td>
<td>0.050</td>
<td>0.040</td>
<td>0.070</td>
<td>1000</td>
<td>0.050</td>
</tr>
<tr>
<td>Tev</td>
<td>1000</td>
<td></td>
<td>0.019</td>
<td>0.012</td>
<td>0.030</td>
<td>1000</td>
<td>0.024</td>
</tr>
<tr>
<td>Tevperm</td>
<td>1000</td>
<td></td>
<td>0.041</td>
<td>0.030</td>
<td>0.055</td>
<td>1000</td>
<td>0.042</td>
</tr>
<tr>
<td>Tuv</td>
<td>1000</td>
<td></td>
<td>0.020</td>
<td>0.013</td>
<td>0.031</td>
<td>1000</td>
<td>0.022</td>
</tr>
<tr>
<td>Tuvperm</td>
<td>1000</td>
<td></td>
<td>0.041</td>
<td>0.030</td>
<td>0.055</td>
<td>1000</td>
<td>0.041</td>
</tr>
<tr>
<td>LRT</td>
<td>1000</td>
<td></td>
<td>0.055</td>
<td>0.042</td>
<td>0.071</td>
<td>1000</td>
<td>0.055</td>
</tr>
<tr>
<td>EST</td>
<td>838</td>
<td></td>
<td>0.340</td>
<td>0.310</td>
<td>0.380</td>
<td>838</td>
<td>0.340</td>
</tr>
<tr>
<td>ESTperm</td>
<td>838</td>
<td></td>
<td>0.050</td>
<td>0.040</td>
<td>0.070</td>
<td>838</td>
<td>0.050</td>
</tr>
<tr>
<td>EST</td>
<td>1000</td>
<td></td>
<td>0.400</td>
<td>0.370</td>
<td>0.430</td>
<td>1000</td>
<td>0.400</td>
</tr>
<tr>
<td>ESTperm</td>
<td>1000</td>
<td></td>
<td>0.060</td>
<td>0.040</td>
<td>0.070</td>
<td>1000</td>
<td>0.050</td>
</tr>
<tr>
<td>Tev</td>
<td>1000</td>
<td></td>
<td>0.026</td>
<td>0.017</td>
<td>0.038</td>
<td>1000</td>
<td>0.026</td>
</tr>
<tr>
<td>Tevperm</td>
<td>1000</td>
<td></td>
<td>0.044</td>
<td>0.033</td>
<td>0.059</td>
<td>1000</td>
<td>0.043</td>
</tr>
<tr>
<td>Tuv</td>
<td>1000</td>
<td></td>
<td>0.026</td>
<td>0.017</td>
<td>0.038</td>
<td>1000</td>
<td>0.026</td>
</tr>
<tr>
<td>Tuvperm</td>
<td>1000</td>
<td></td>
<td>0.044</td>
<td>0.033</td>
<td>0.059</td>
<td>1000</td>
<td>0.043</td>
</tr>
</tbody>
</table>
Table B.4: Power estimates and 95% confidence intervals calculated for SAMP and LINK α-samples under recessive inheritance. Reps is the number of replicates performed in the presence of heterogeneity.

<table>
<thead>
<tr>
<th>Reps</th>
<th>SAMP</th>
<th>Reps</th>
<th>LINK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>95% CI</td>
<td>Power</td>
</tr>
<tr>
<td>0.8</td>
<td>LRT</td>
<td>500</td>
<td>0.904</td>
</tr>
<tr>
<td>0.8</td>
<td>LRT*</td>
<td>500</td>
<td>0.934</td>
</tr>
<tr>
<td>0.8</td>
<td>EST1</td>
<td>479</td>
<td>0.960</td>
</tr>
<tr>
<td>0.8</td>
<td>EST1perm</td>
<td>479</td>
<td>0.960</td>
</tr>
<tr>
<td>0.8</td>
<td>EST1*</td>
<td>479</td>
<td>0.511</td>
</tr>
<tr>
<td>0.8</td>
<td>EST</td>
<td>500</td>
<td>0.990</td>
</tr>
<tr>
<td>0.6</td>
<td>ESTperm</td>
<td>500</td>
<td>0.960</td>
</tr>
<tr>
<td>0.6</td>
<td>EST*</td>
<td>500</td>
<td>0.402</td>
</tr>
<tr>
<td>0.6</td>
<td>Tev</td>
<td>500</td>
<td>0.988</td>
</tr>
<tr>
<td>0.6</td>
<td>Tevperm</td>
<td>500</td>
<td>0.988</td>
</tr>
<tr>
<td>0.6</td>
<td>Tev*</td>
<td>500</td>
<td>0.996</td>
</tr>
<tr>
<td>0.6</td>
<td>Tuv</td>
<td>500</td>
<td>0.988</td>
</tr>
<tr>
<td>0.6</td>
<td>Tuvperm</td>
<td>500</td>
<td>0.988</td>
</tr>
<tr>
<td>0.6</td>
<td>Tuv*</td>
<td>500</td>
<td>0.996</td>
</tr>
<tr>
<td>(a)</td>
<td>Test</td>
<td>Reps</td>
<td>SAMP</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>0.5</td>
<td>LRT</td>
<td>500</td>
<td>0.992</td>
</tr>
<tr>
<td>0.5</td>
<td>LRT*</td>
<td>500</td>
<td>0.992</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1</td>
<td>500</td>
<td>0.990</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1perm</td>
<td>500</td>
<td>0.980</td>
</tr>
<tr>
<td>0.5</td>
<td>EST1*</td>
<td>500</td>
<td>0.588</td>
</tr>
<tr>
<td>0.5</td>
<td>EST</td>
<td>500</td>
<td>0.990</td>
</tr>
<tr>
<td>0.5</td>
<td>ESTperm</td>
<td>500</td>
<td>0.980</td>
</tr>
<tr>
<td>0.5</td>
<td>EST*</td>
<td>500</td>
<td>0.496</td>
</tr>
<tr>
<td>0.5</td>
<td>Tev</td>
<td>500</td>
<td>0.992</td>
</tr>
<tr>
<td>0.5</td>
<td>Tevperm</td>
<td>500</td>
<td>0.992</td>
</tr>
<tr>
<td>0.5</td>
<td>Tev*</td>
<td>500</td>
<td>0.996</td>
</tr>
<tr>
<td>0.5</td>
<td>Tuv</td>
<td>500</td>
<td>0.992</td>
</tr>
<tr>
<td>0.5</td>
<td>Tuvperm</td>
<td>500</td>
<td>0.992</td>
</tr>
<tr>
<td>0.5</td>
<td>Tuv*</td>
<td>500</td>
<td>0.996</td>
</tr>
</tbody>
</table>


