Statistical Power Analysis and Related Issues in Human Genetic Linkage and Association

by

Jing Hua Zhao

Bachelor of Medicine, Shandong Medical University (1985)

Master of Medicine, Shanghai Medical University (1988)

A thesis submitted for the degree of Doctor of Philosphy of the University of London

Division of Psychological Medicine

Institute of Psychiatry

King's College London

The United Kingdom

May 2004

TO MY FAMILY

Statistical Power Analysis and Related Issues in Human Genetic Linkage and Association

by

Jing Hua Zhao

Bachelor of Medicine, Shandong Medical University (1985)
Master of Medicine, Shanghai Medical University (1988)

Submitted to the King's College London
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

University of London

May 2004

Author .....
Division of Psychological Medicine
May, 2004

Certified by .....
Pak Chung Sham
Professor
Thesis Supervisor

Accepted by .....
Faculty of Medicine
Board of Studies Genetics

Statistical Power Analysis and Related Issues in Human Genetic Linkage and Association

by

Jing Hua Zhao

Submitted to the King's College London
in May, 2004, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The difficulty with genetic study of complex traits has raised concerns over optimal study design and statistical analysis. The key statistical issues in designing different studies are validity, power and robustness of relevant statistical tests. This thesis investigates several scenarios of power analysis in linkage and association analysis, which include linkage tests in small pedigrees, association tests for case-control data, marker polymorphism and mutation detection, computer simulation methods and application. It also gives results of numerical experiments on family haplotype analysis and discusses TDT and other association designs. This thesis reveals that commonly used parametric and nonparametric linkage statistics are comparable in power for two point analysis with simple families, but nonparametric linkage statistics are anticonservative in some families. As for case-control data, heterogeneity statistic nearly has power close to the true model without the needs of disease model specification, and comparable to the ordinary likelihood ratio test for contingency table. In mutation detection, multiallelic marker is usually more favourable than SNP. While haplotype analysis has claimed power for linkage and association, there may be numerical and analytical difficulties when a likelihood approach using family data is adopted. Finally, correct sample sizes are obtained for TDT design as reported earlier in the literature, and some computer routines performing these calculations are also given.

Acknowledgments

I would like to thank Professor Pak Sham and Dr David Curtis for their support and supervision, which makes it possible for me to conduct this project. I am also grateful of Professor Robin Murray for his support for the final year of my work.

As a six-year tenant in Institute of Psychiatry I am indebted to many people. I wish to express special thanks to the former personnel manager Ms Catriona Urquart for efficiently sorting out my visa several times, to Jenny Merchant, Andrew Wallis, Muriel Walsh and Maureen Armstrong-Smith for their help with departmental affairs, to Lee Wilding and Debbie Heavey for academic information, to Malcolm Hart and Eric Glover for their help with computing, to Professor Peter McGuffin for reading and supporting my upgrade report, and to Ms Marion Cuddy for reading through final version of the draft, and to all friends in the institute for their help on many other matters. Chapters 1-4 of he thesis were improved at suggestions from Professors John Whittaker (Imperical College, London) and Tim Bishop (Cancer Research UK, St James's University Hospital, Leeds). Occasional revisions to Chapter 2 of the thesis were made while working as a statistician in the Department of Epidemiology and Public Health, University College London over a period of about two years, but due to my new role in the Whitehall II study including more commitment to longitudinal data analysis and social epidemiology, no more references have been added since my viva in July 2002. I am grateful of Professor Sir Michael Marmot and his Whitehall II team for making this possible.

I am grateful of Professors Zhaohuan Zhang and Fumin Shen of Shanghai (Fudan) Medical University for igniting my interests in genetic epidemiology during my master training between 1985 and 1988, Professor Xiping Xu of Harvard University for resuming my work in this area from 1995, Professors Jin Ma and Yuanli Liu of Harvard University, Professor Pihuan Jin of Shanghai Medical University for supporting my coming to the UK. Thanks also to my former colleagues and friends in China for their concern over so many years.

Finally I would like to thank my special friend, Dr Wendi Qian, for reading through my upgrade report as well as the first draft of this manuscript and giving many helpful comments. I wish to dedicate this thesis to my family for their patience, love and care.

Table of Contents

1 Preliminaries **1**

1.1 Basic terminology 1

1.2 Mean and variance of noncentral χ^2 distribution 5

1.3 Segregation analysis 6

1.4 Linkage analysis 9

1.5 Association analysis 15

1.6 Coalescent models 19

1.7 Outline of thesis 22

1.8 Bibliographic notes 22

2 Linkage tests for small pedigrees **24**

2.1 Introduction 24

2.2 Methods 25

2.3 Results 33

2.4 Discussion 35

2.5 Bibliographic notes 37

3 Case-control allelic association **38**

3.1 Introduction 38

3.2 Methods 39

3.3 Results 43

3.4 Discussion 45

3.5 Bibliographic notes 46

4 Marker characteristics and fine mapping **49**

4.1 Introduction 49

4.2 Methods 50

4.3 Results 52

4.4 Discussion 56

4.5 Bibliographic notes 57

5 Computer simulation methods **60**

5.1 Introduction 60

5.2 Methods, implementations and results 61

5.3 Discussion 72

5.4 Bibliographic notes 74

6 Association in nuclear families **76**

6.1 Introduction 76

6.2 Methods 77

6.3 Results 81

6.4 Discussion 82

6.5 Bibliographic notes 85

7 Transmission/disequilibrium tests **86**

7.1 Introduction 86

7.2 Methods 87

7.3 Results 89

7.4 Discussion 91

7.5 Bibliographic notes 92

8 Bells and Whistles **95**

8.1 Major findings 95

8.2 Possible extensions 96

8.3 Epilogue 97

A Gene counting method **98**

B Computer programs **114**

B.1 C subroutines 114

B.2 SAS programs 127

B.3 QBASIC program 132

List of Tables

1.1 Relationship between parameters in the Mendelian "mixed models" 8

2.1 The four genetic models 28
2.2 Estimated sample sizes required under linkage homogeneity 33
2.3 Estimated sample sizes required under linkage heterogeneity 34
2.4 Proportion of replicates (out of 10,000) simulated under the null hypothesis reaching threshold of 0.001 35

3.1 Description of test statistics 40
3.2 The five genetic models 42
3.3 Mean and standard deviation of χ^2 statistics from 1,000 subjects 43
3.4 Mean and standard deviation of χ^2 statistics from 10,000 subjects 44
3.5 Estimated sample sizes required for 90% power and significance level .05 . . 44

4.1 Required sample sizes for $\alpha = 0.001$ and $1 - \beta = 0.9$ over $g = 20, 50, 80$ and $\theta = 0.005, 0.05$ 52

5.1 Some model constants 62

6.1 The 22 models compatible with Alzheimer's 79

7.1 Comparison of linkage and association in nuclear families required for identification of disease gene 90
7.2 Estimated sample sizes required for association detection 91

A.1 Symbols and their meanings 98
A.2 Genotype counts for biallelic markers 99
A.3 Genotypic probabilities for two biallelic markers 99

A.4 Genotype count for three biallelic markers 100
A.5 The genotypic probabilities for three biallelic markers 101
A.6 Initial count table for disease locus and a marker 102
A.7 Genotypic probabilities for a disease locus and a marker 102
A.8 The updated count table for a disease locus and a marker 102
A.9 Initial count table for a disease locus and two markers 103
A.10 The genotypic probabilities for a disease locus and two markers 104
A.11 The updated gene count table for a disease locus and two markers 105
A.12 Gene counts for a triallelic disease locus and two biallelic markers 106

List of Figures

2-1 The four family types 30

5-1 Flowchart of SIM 63
5-2 Familial hemophagocylic lymphohistiocytosis family 1 65
5-3 Familial hemophagocylic lymphohistiocytosis family 2 66
5-4 Familial hemophagocylic lymphohistiocytosis family 3 67
5-5 Familial hemophagocylic lymphohistiocytosis family 4 68
5-6 Familial hemophagocylic lymphohistiocytosis family 5 69
5-7 The Leukemia family 71

6-1 Pedigree for association analysis between two marker loci 78
6-2 LOD score of 3S family assuming linkage equilibrium 83
6-3 LOD score of 3S family assuming linkage disequilibrium 84

Abbreviations

AFBAC	Affected family based-controls
ASP/APM	Affected sib pair/affected pedigree member
cM	CentiMorgan
DNA	deoxyribonucleic acid
ELOD	Expected lod score
EM/ECM	Expectation/Conditional-maximisation
FFT	Fast Fourier Transform
GAW	Genetic analysis workshop
GEE	Generalised estimating equations
HWE	Hardy-Weinberg equilibrium
IBD/IBS	Identity by descent/state
IML	Interactive matrix language (SAS)
LD	Linkage disequilibrium
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimate
MLS	Maximum lod score
MRCA	Most recent common ancestor
NCP	Noncentrality parameter
NPL	Nonparametric linkage statistic
QTL	Quantitative trait locus
RFLP	Restricted-fragment length polymorphism
SNP	Single nucleotide polymorphism
TDT	Transmission/disequilibrium test

References to computer programs

APM	Weeks & Lange (1988)
ASPEX	Hinds & Risch (1996)
CHRSIM	Speer et al. (1992), Terwilliger et al. (1993)
EH/EHPLUS	Xie & Ott (1993), Zhao J et al. (2000)
ESPA	Sandkuijl (1989)
FASTLINK	Cottingham et al. (1993)
GASP	Wilson et al. (1996)
GENEHUNTER	Kruglyak et al. (1996)
GENEHUNTER PLUS	Kong & Cox (1997)
HAPLO	Hawley & Kidd (1995)
LIPED	Ott (1974)
LINKAGE	Lathrop et al. (1994)
MENDEL	Lange et al. (1988)
MFLINK	Curtis & Sham (1995)
MORGAN	Thompson (2001)
MULTTDT	Zhao et al. (2000)
PAP	Hasstedt & Cartwright (1981), Hasstedt (1982, 1994)
PDT	Martin et al. (2000)
POINTER	Lalouel & Morton (1981), Morton et al. (1983)
SIMLINK	Ploughman & Boehnke (1989)
SLINK	Ott (1989), Weeks et al. (1990)
SOLAR	Almasy & Blangero (1998)
SPLINK	Holmans (1993)
TRANSMIT	Clayton (1999)
VITESSE	O'Connell & Weeks (1995), O'Connell (2001)

Chapter 1

Preliminaries

1.1 Basic terminology

There is a long history of interests in heritable characters. In 1805, Gregor Mendel discovered his laws of inheritance. Later, Francis Galton and his student Karl Pearson observed that family resemblance of many traits did not show Mendelian patterns of inheritance, but that trait value of offspring tended to be, on average, midway between the parents, with some variability. In 1918, Ronald Fisher showed that continuous trait may be determined by small effects of many genes. By central limit theorem, a large number of such effects would result in a normal distribution. The concept of the gene was finally materialised in 1953 by Watson and Crick who unveiled the duplex structure of DNA (deoxyribonucleic acid). Until early 1980s, the study of association between genetic markers and human disorders was much limited by the number of genetic markers available, but this has been rapidly changed by the Human Genome Project, which has created an unprecedented opportunity and multidisciplinary collaboration including biologists, medical doctors, mathematicians and computer scientists.

The total amount of genetic information in individual possesses is called the genome. Gene is the basic unit of inheritance. Human genes are stored in 23 pairs of chromosomes – 22 pairs of autosomes and 1 pair of sex chromosomes (XX for females or XY for males) which are found in the cell nucleus. The position at a chromosome where a gene is located is called **locus**. A gene may have two or more variant forms, each called **allele**, and such a gene is called biallelic or multiallelic. An individual has two alleles (which may be the same or different) for each gene, one on each of the two homologous chromosomes,

at the appropriate locus for that gene. At the molecular level, genes are specific DNA coding sequences. There are approximately 3×10^9 nucleotides in the human genome. The lengths of human genes are variable (Eyre-Walker & Keightley 1999), and according to the estimate that a typical genes is about 30,000 nucleotides long, these nucleotides form about 50,000$\sim$100,000 genes (Cantor 1994; Thompson 1996; Lange 1997; Ott 1999). This estimate has been recently been refined to about 30,000. Although this estimate is based on several sources, only some of them overlap (Hogenesch et al. 2001). In addition to nucleic DNA there is also mitochondrial DNA (mtDNA). The mitochondrion is a small haploid nonnucleic DNA organelle about 17,000 base pairs long, which is primarily maternally inherited.

A lot of current biomedical researches involves the characterisation of genes and their functional roles, much of this research is possible due to availability of **genetic markers**, i.e., anonymous sequences of DNA that shows polymorphic variation, such as restricted fragment-length polymorphism (RFLP), simple sequence repeat (SSR), and single nucleotide polymorphism (SNP). The **human genome project** is a global effort towards understanding the human genome and includes development of the following tools (Lander & Waterman 1995): **genetic maps**, to produce a genetic map showing the location of 5,000 polymorphisms that can be used to trace inheritance of diseases in families; **physical maps**, to produce a collection of overlapping pieces of DNA that cover all the human chromosomes; and **DNA sequence**, to sequence the entire genome. In early 2001, the first complete draft of human genome was produced. As of October 2001, a total of 9,588 genes or phenotypes had been established (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov).

Simple traits such as cystic fibrosis, Huntington disease and diastrophic dysplasia are controlled by a single gene and called **monogenic**, as compared to **oligogenic** traits where several genes are involved. Many genes control some part of the developmental process, and are sometimes referred to as 'switch genes'. The maternal and paternal alleles of an individual at a locus define the **genotype** of the individual at that locus. Alleles (at different loci) received from one parent are referred to as **haplotype**. The **phenotype** is an observable or measurable trait of an individual. The relationship between genotype and phenotype can be characterised by **penetrance**, the conditional probability of observing a phenotype given a genotype.

Genetic analysis often involves **pedigree**, which is a collection of individuals who are genetically related. Individuals in a pedigree are either **founders** whose parents are not in

the pedigree or **non-founders** whose parents are in the pedigree. A large pedigree may consist of many **nuclear families** each including parents and their offsprings (children). **Mendel's first law** states that for any gene, each parent transmits one allele chosen at random to its offspring. **Mendel's second law** declares that for any two genes, the alleles transmitted by a parent are independent. The separation of the two members of a pair of alleles during the formation of gametes is called **segregation**, and is due to the separation of the two homologous chromosomes during meiosis. In fact, genes on the same chromosomes tend to be inherited together, a phenomenon called **linkage**, so the second law generally turns out to be true only for non-syntenic loci. In meiosis, each member of a pair of homologous chromosomes replicates to form two sister chromosomes known as chromatids. The maternally and paternally derived chromatids cross over to form chromosome of offspring. **Crossing-over** is the reciprocal exchange of segments between non-sister chromatids during meiosis. A recombination is said to have occurred between two loci, if alleles at those loci on a single offspring chromosome are of different grand-parental origin. **Interference** is the lack of independence in recombinations at different intervals on a chromosome.

The risk for developing many common diseases such as diabetes is influenced by both genetic and environmental factors. These diseases are called **complex traits** because of their complex aetiology (Lander & Schork 1994). Complexity may take the form of: **incomplete penetrance**, the probability that individuals inheriting the gene will have the disease phenotype may be less than 1 and may depend on age or other factors; **phenocopy**, a disease can be due to nongenetic causes; **genetic heterogeneity**, a disease may be due to different genetic mutations in different individuals; and **polygenic inheritance**, the liability of disease is due to the additive or interactive effects at multiple loci. There are other phenomena such as **imprinting**, the difference in function of an allele according to its parental origin. Statistically, the term **additive** refers to the accumulative effects of individual alleles, **dominance** refers to the interaction between alleles at the same locus, and **epistasis** refers to the interaction between alleles at different loci. **Pleiotropic expression** is a situation in which an allele has more than one distinct phenotypic effect.

Development of statistical and computational methods can be classified into two broad categories: the inference of underlying genetic mechanism assuming the relationship structure of the observed individuals or population is known (in both experimental and natu-

ral populations), and the inference of structure assuming the underlying biological basis is known (usually human studies). In experiment where individuals of an organism can be chosen for breeding, design is a major issue, and missing data are few. In human genetic studies, problems of missing data and sampling procedures are paramount (Thompson 1996). While the data necessary to understand different complex traits is disease-specific, the underlying principles used to identify and dissect the genetic component are applicable to a variety of diseases. The **positional cloning** approach is to map and identify previously unknown loci, which involves two initial types of studies. First, a disease model is developed to approximate the mode of inheritance of the disease. Second, one or more of the relevant genes is localised (mapped) to a chromosomal region by a systematic screening of the whole genome. The **candidate gene** approach is to study loci which are suspected to play a key role in the disease. There are two major difficulties in genetic study of human diseases: (1) for human diseases, one cannot arrange matings at will but rather must retrospectively interpret existing families; and (2) for human diseases the trait may not be simply related to the genotype at a single gene (although this may be true also for many traits in experimental organism). Even with a dense genetic map of DNA polymorphisms, human genetic mapping confronts several special problems of incomplete information (Lander 1995): for individuals homozygous at a gene, one cannot distinguish between the two homologous chromosomes based on the genotype at the locus; for individuals heterozygous at a gene, one cannot tell which allele is on the paternal chromosome and which is on the maternal chromosome unless one can study the individual's parents; and information for deceased individuals (or for those who choose not to participate in a genetic study) is completely missing. For these reasons often one cannot simply count recombinants directly to estimate recombination frequencies.

The theory of **population genetics** is increasingly used in human genetics (e.g. Schork et al. 1998; Weeks & Lathrop 1995). For each generation there is an element of chance in the drawing of gametes that will unite with other to form the next generation. Chance alone can result in changes in allele frequency, and as this change does not follow any predetermined direction, it is known as random genetic drift (Hartl & Clark 1997). Mutation is a change in DNA sequence and provides new variation. Migration among populations or shared selection decreases divergence among them. Selection (e.g., differential survival or fertility of genes of certain allelic types) in favour of a specific allele may increase frequency of the

haplotype containing that allele. Assuming random mating and absence of disturbing forces such as migration, mutation, and selection for the gene of interest, the genotype distribution is governed by Hardy-Weinberg equilibrium (HWE). Therefore for a gene with two alleles A and a, each with frequency p and q, the frequencies of genotypes AA, Aa and aa would be p^2, $2pq$ and q^2. Under random mating the genotype distribution remains in these Hardy-Weinberg proportions from one generation to the next. **Assortative mating** refers to any system of mating in which mate choice is influenced by genotype or phenotype. In positive assortative mating, mates are more like each other than they would be by chance, and such mating tends to increase homozygosity.

To make inferences from genetic data, a statistical model is often required. Essentially, **statistical inference** is concerned with parameter estimation and hypothesis testing. The optimality of a statistical test is can be assessed by many criteria: the **validity** of a test indicates if it produces the correct type I error under the null hypothesis; the probability of rejecting the null hypothesis when an alternative is true is **power**; the resistance to departures from the idealised assumptions is called **robustness**.

1.2 Mean and variance of noncentral χ^2 distribution

Since statistical inference is often via likelihood ratio tests which are asymptotically χ^2, noncentral χ^2 distribution is often used in statistical power analysis. Here we give the characteristic function, cumulant generating function and conditional expectations. The characteristic function for a random variable ξ with probability measure P is the complex valued function (Fourier transform) of the variable $t \in R^1$ given by $\phi(t) = E(\exp(itx)) = \int_{-\infty}^{\infty} \exp(itx) dP(x)$ where $i = \sqrt{-1}$. If ξ has an absolute moment of order k, then $\phi(t)$ is k times continuously differentiable and $\phi^{(k)}(0) = i^k E\xi^k, k = 0, 1, \ldots$ (Note that the definition and result apply to both discrete and continuous random variables). Now let X_i, $i = 1, \ldots, \nu$ be independent variables with distribution and $N(\mu_i, 1)$ characteristic function $\exp(\mu_i^2 it/(1-2it))/\sqrt{1-2it}$ (Grimmett & Stirzaker 1992), then random variable $Y = \sum_{i=1}^{\nu} X_i^2$ has characteristic function $\phi(t) = \exp[(it\delta)/(1-2it)]/[(1-2it)^{\nu/2}]$, where $\delta = \sum i = 1^{\nu} \mu_i^2$. The random variable Y is said to have a noncentral χ^2 distribution with ν degrees of freedom and noncentrality parameter (NCP) δ ($\chi^2_{(\nu,\delta)}$). Further we have $\phi^{(1)}(t) = \phi(t)(i\nu + 2t\nu + i\delta)/(1-2it)^2$, $\phi^{(2)}(t) = \phi^{(1)}(t)[(4i + i\nu)(1-2it) + i\delta]/(1-2it)^2$

$+ (2\nu)/(i\nu + 2t\nu + i\delta)$ so $E(Y) = \nu + \delta$, $E(Y^2) = -\phi^{(2)}(0) = (\nu + \delta)(4 + \nu + \delta) - 2\nu$ and $V(Y) = 2\nu + 4\delta$. Similar expressions for moments of arbitrary order can be obtained from $\log(E(\exp(tx)))$ called cumulant. The rth cumulant is $k_r = 2^{r-1}(r-1)!(\nu + r\delta)$ (Patnaik 1949). The mean and variance can also be derived in a hierarchical model setting (Casella & Berger 1990, p157), $Y|K \sim \chi^2_{n+2k}, K \sim Poisson(\delta/2)$ then $E(Y) = E(E(Y|K)) = E(\nu + 2K) = \nu + \delta$, similarly $V(Y) = V(E(Y|K)) + E(V(Y|K)) = V(\nu + 2K) + E(2(\nu + 2K)) = 4V(K) + 2\nu + 4E(K) = 2\nu + 4\delta$. We have used the fact that $E(K) = V(K) = \delta/2$. Power of a χ^2 statistic given significance level α can be obtained as $\int_{\chi_{\alpha(\nu,0)}}^{\infty} d\chi^2(\nu,\delta)$, where $\chi^2_{\alpha(\nu,0)}$ is the $100(1-\alpha)$ percentage point of the central χ^2 with ν degrees of freedom, and δ is the NCP.

1.3 Segregation analysis

The inference of mode of inheritance from pedigree data is called segregation analysis. Complex segregation analysis (Elandt-Johnson 1971; Lalouel & Morton 1981) is the statistical method used in handling family data when two or more distinct and functionally independent segregation parameters are involved in the analysis. Traditionally it deals with the following situations (Elandt-Johnson 1971, p497) "*(a) There are two phenotypic forms of a certain genetic character; neither form is rare, and both are controlled by one (or more) loci. (b) There are more than two phenotypic forms controlled by a certain genetic system (e.g. multiallelic inheritance or multilocus inheritance with or without epistasis and/of with or without linkage). (c) Some of the models discussed in (a) or (b) can be complicated by incomplete penetrance or different fitnesses of genotypes*".

The likelihood of phenotypic data X of a pedigree with size n can be written as follows

$$L(X) = \sum_{g_1}\sum_{g_2}\cdots\sum_{g_n} \Pi_{j=1}^{n} P(X_j|g_j)\Pi_{k=1}^{f} P(g_k)\Pi_{m=1}^{n_f}\tau(g_m|g_{m1}, g_{m2}) \qquad (1.1)$$

where g_i is the multilocus genotype of individual i, f the number of founders, n_f the number of nonfounders, $P(.)$ the genotypic frequency, $P(.|.)$ the penetrance, $\tau(.|.,.)$ the transmission function, m_1 and m_2 the parental indices of individual m. The tranmission function (Elston & Stewart 1971) specifies the probabilities of the three major locus genotypes (AA, Aa, aa) transmitting the normal allele (A), to be (1,0.5,0) in the Mendelian case, and test of heterozygote transmission probability being 0.5 may detect departure from Mendelian

model, while setting the transmission probabilities to be the same tests the transmission of major determinants (Morton et al. 1983). An algorithm for calculation of expression (1.1) in general pedigree was developed by Elston and Stewart (1971). Expression (1.1) can be extended to incorporate environmental factors or covariates. Given family data, likelihoods can be obtained under models ranging from Mendelian, non-Mendelian to the general models, from which the best-fitting model is selected (Elandt-Johnson 1970; Rao & Elandt-Johnson 1971; Morton & MacLean 1974; Boyle & Elston 1979; Elston 1980; Morton et al. 1983).

Under a generalised single locus model in which the disease locus is biallelic with disease allele a (with frequency q) and normal allele A (with frequency $1-q$), there are three possible genotypes AA, Aa and aa. The mode of inheritance can be delineated by the frequency of a and penetrances for AA, Aa and aa. Therefore Mendelian dominant trait has mode of inheritance (q, 0, 1, 1) and Mendelian recessive trait as (q, 0, 0, 1), while non-Mendelian trait has penetrances taking any values between 0 and 1. Under a mixed model, a phenotype of interest (x) is more appropriately assumed to be a result of major locus, polygenic, common and random environmental effects (Morton & MacLean 1974). Suppose the major locus is biallelic with alleles A and a, genotypic effects of AA, Aa and aa can be characterised by z, q, t, d, where z is the mean effect of AA, q is the allele frequency of a, t is the displacement between AA and aa and d is the dominance. The genetic effects of genotypes AA, Aa and aa are z, $z+dt$ and $z+t$ with frequencies p^2, $2pq$, and q^2 under HWE. When the major locus effect has mean 0 and variance 1, we have $z=-(tq^2+2pqdt)$ and $t^2=1/[(pq)^2(q+2pd)^2+2pq(d-q^2-2pqd)^2+q^2(1-q^2-2pqd)^2]$, so that the only q and d are free parameters. Computer program **POINTER** (Morton & MacLean, 1974; Lalouel & Morton, 1981; Morton et al. 1983) incorporates male and female mutations and X-linked case, and obtains maximum-likelihood estimates (MLEs) by iterating upon the following: μ, mean of x; V, variance of x; d, t, q, as above; H, polygenic heritability; and B, relative variance due to common environment. Phenotype specification may concern a quantitative measurement, a dichotomous affection status, or both (Morton et al. 1983). Phenotypic liability to affection y, is related to x by $y=x+w$, where $w \sim N(0,W)$ and $cov(x,w)=0$. Individuals are affected whenever liability x is greater than some threshold. For such a specification to be independent of any particular model of inheritance in segregation analysis, one must provide an estimate of the prior probability of affection

in the reference population. **POINTER** was extended by **PAP** (Hasstedt & Cartwright 1981; Hasstedt 1982, 1994) which parameterises total mean (μ), the total standard deviation (σ), the frequency (q), dominance (d) and displacement (t) at major locus and polygenic heritability (h^2), and parent-to-offspring transmission probabilities (τ_1, τ_2, τ_3) for the three genotypes at one locus. Major loci could be inferred sequentially, i.e., after one major locus is inferred it is fixed with its estimated parameters and the second major locus is tested, and so on (Hasstedt et al, 1997). The relationship between **POINTER** and **PAP** can be summarised in Table 1.1 (Khoury et al. 1993; Ziegler & Hebebrand 1998).

Table 1.1: Relationship between parameters in the Mendelian "mixed models"

PAP	**POINTER**	Conversion
p frequency of A_1 allele	q frequency of low allele	note if μ_1 is "low"
μ_1 mean of genotype A_1A_1	μ population mean	$\mu = p^2\mu_1 + 2pq\mu_2 + q^2\mu_3$
μ_2 mean of genotype A_1A_2	V population variance	$V = \sigma^2_W + \sigma^2_{ML}$‡
μ_3 mean of genotype A_2A_2	t displacement between homozygotes	$t = \mu_3 - \mu_1$
σ^2_W within genotype variance	d deviation due to dominance	$d = \mu_2 - \left(\frac{\mu_3 - \mu_1}{2}\right)$
H_p proportion of σ^2_W attributed to additive polygenes=(σ^2_A/σ^2_W)	h^2 narrow sense heritability=$\frac{\sigma^2_A}{V}$	$h^2 = H_p\left(\frac{\sigma^2_W}{\sigma^2_W + \sigma^2_{ML}}\right)$

‡Variance due to major locus, $\sigma^2_{ML} = p^2(\mu_1 - \mu)^2 + 2pq(\mu_2 - \mu)^2 + q^2(\mu_3 - \mu)^2$

Assuming Mendelian inheritance, genetic effects on a trait was traditionally examined for multi-modalities of its distribution (e.g. MacLean et al. 1976). Genetic covariate effects on continuous and discrete traits can be examined via regressive models (Bonney 1984, 1986) and implemented in **SAGE** (http://darwin.cwru.edu) modules REGD, REGC and REGTL. Methods using generalised estimating equation (GEE) were also proposed (e.g. Lee et al. 1993; Zhao 1994; Lee & Stram 1996; Trégouët & Tiret 2000).

Owing to ascertainment in affected individuals, segregation analysis may suffer from bias if an ascertainment correction is not made (Ewens 1991). The power of segregation analysis was investigated by Konigsberg et al. (1989), Borecki et al. (1994, 1995). Genetic marker can in principle be incorporated into a formal combined segregation and linkage analysis (e.g. Borecki et al. 1994). Such models have also been extended to consider polygenic and multifactorial effects.

1.4 Linkage analysis

The study of cosegregation of genetic markers with putative disease genes in families is called **linkage analysis**. Alternatively, it "*refers to the ordering of genetic loci on a chromosome and to estimating genetic distances among them*" (Ott 1999). It requires specification of the relationship between map distance and recombination. The simplest case assumes there is no interference so crossovers occur as a Poisson process. In this case, map function $M(x)$ is obtained from the probability that a Poisson random variable with mean x is odd, that is, $\sum_{k=odd} \exp(-x)x^k/k! = \exp(-x)/2 \sum_k \left(x^k/k! - (-x)^k/k!\right) = (1 - \exp(-2x))/2$. It can also be derived from $M(x+\delta x) = M(x)(1-\delta x) + (1 - M(x))\delta x$, $dM(x)/dx = (1 - 2M(x))$, $\ln(1 - 2M(x)) = -2x$, $M(x) = (1 - \exp(-2x))/2$ since $M(0) = 0$. Recent discussions of models for recombination include Zhao et al. (1995a, 1995b), Lange (1997), Speed (1997), Ott (1999), Broman & Weber (2000), and Thompson (2000c).

Equation (1.1) underlies the Elston-Stewart method for linkage analysis (Elston & Stewart 1971; Cannings et al. 1978). Methods based on hidden Markov model (Baum et al. 1970, 1972), was developed by Lander & Green (1987). The Elston-Stewart algorithm is linear on number of individuals while the Lander-Green algorithm is linear on the number of markers. The Lander-Green algorithm is based on the notion of inheritance vectors, indicator arrays with element taking values 0 or 1 to indicate whether paternal or maternal allele is transmitted. Consider a pedigree with N meiosis and three loci A, B and C with inheritance vectors α, β and γ, the probability of observing phenotype data can be expressed in terms of these inheritance vectors. For loci A and B, the probability of observing phenotype data a and b can be expressed as follows (Idury & Elston 1997)

$$\begin{aligned} p(a,b) &= \sum_{\alpha,\beta} p(a,b|\alpha,\beta)p(\alpha,\beta) \\ &= \sum_{\alpha,\beta} p(\alpha,\beta)p(a|\alpha)p(b|\beta) \\ &= \sum_{\alpha} p(\alpha)p(a|\alpha) \sum_{\beta} p(\beta|\alpha)p(b|\beta) \\ &= 2^{-N} P_a T_1 P_b \end{aligned} \tag{1.2}$$

where T_1 is $2^N \times 2^N$ transmission matrix with elements $t_{\alpha,\beta} = \theta_1^p(1-\theta_1)^{N-p}$, θ_1 is the reombination rate, p being determined by α and β. When the third locus C is included we

have.

$$\begin{aligned} p(a,b,c) &= \sum_{\alpha,\beta,\gamma} p(a,b,c|\alpha,\beta,\gamma)p(\alpha,\beta,\gamma) \\ &= \sum_{\alpha,\beta,\gamma} p(\alpha,\beta,\gamma)p(a|\alpha)p(b|\beta)p(c|\gamma)p(\gamma|\beta,\alpha)p(\beta|\alpha)p(\alpha) \\ &= \sum_{\alpha} p(\alpha)p(a|\alpha) \sum_{\beta} p(\beta|\alpha)p(b|\beta) \sum_{\gamma} p(\gamma|\beta)p(c|\gamma) \\ &= 2^{-N} P_a T_1 P_b T_2 P_c \end{aligned} \tag{1.3}$$

where T_2 is defined similarly to T_1. This yields an algorithm faster than the original Lander-Green algorithm. Kruglyak & Lander (1998) made further improvement to (1.2) and (1.3) by fast Fourier transform (FFT). Strauch et al. (2000) extended it to incorporate imprinting and two-locus model. With the limitations of both Elston-Stewart and Lander-Green algorithms, Markov chain Monte Carlo (MCMC) methods and associates become important alternatives (Lange & Sobel 1991; Thomas & Cortessis 1992; Thompson et al. 1993; Kong et al. 1993; Thompson 1995; Lin 1996; Heath 1997; Thompson 2000a, 2000c; Lee & Thomas 2000).

The test of linkage is via the ratio of likelihoods assuming linkage and no linkage. The base-10 logarithm of the ratio is termed the lod score (Morton 1955), so that $2\ln 10$ lod is approximately χ^2 distributed with one degree of freedom. The **location score** is the natural log of likelihood raio assuming linkage and no linkage, closer to χ^2 without the need to multiply $\ln 10$. The alternative hypothesis of linkage is restricted to $0 < \theta < 0.5$, therefore the test turns out to be one-tailed, the pointwise significance can be determined by 0.5 ($\chi^2 > 2\ln 10$ lod), since $2\ln 10$ lod is a 50:50 mixture of a point mass at 0 and χ^2_1 (Self & Liang 1987; Nyholt 2000). A lod score of 3, or a likelihood ratio of 1,000, is commonly used to indicate significant linkage, which can be translated into the odds of ~20:1 in favour of linkage, or a significance level of 0.05 (Ott 1991). In the context of genome scanning using pedigree data, the genomewide significance of 0.05 is achieved at a pointwise significance of 4.9×10^{-5}, or lod score of ~ 3.3 (Lander & Kruglyak 1995; Ott 1999; Nyholt 2000). A lod score of less than -2.0 is customarily accepted as conclusive evidence for the exclusion of linkage. Genetic heterogeneity can be allowed for using the admixture model for heterogeneity (Smith, 1963). Under this model, the probability of the trait being linked in a given pedigree is α; with probability $1-\alpha$ the trait is unlinked. This

model assumes that while different pedigrees may have different genetic forms of the disease, within a pedigree only a single genetic form is present. If genetic heterogeneity is allowed for, two different lod scores are calculated: the standard lod score which assumes genetic homogeneity, and a lod score which allows for maximisation of the likelihood function over both the recombination fraction and the linked fraction α. For rare recessive disorders, affected individuals in inbred families tend to have more homozygosity by descent at the disease locus, this leads to the method of homozygosity mapping (Smith 1953; Lander & Bostein 1987).

Power for a traditional linkage study is the probability of obtaining a maximum lod score ≥ 3 for a linked marker (Morton 1955). If the number of recombinants and non-recombinants can be obtained directly (in the so-called direct method of linkage, say offsprings in a family of a phase-known double backcross mating can be scored recombinants and nonrecombinants, Ott 1985), this could be particularly simple. For a given true recombination fraction, θ, significance level α, and a power $1-\beta$ (e.g. $\alpha = 0.0001$, $z_{1-\alpha} = 3.7$; $\beta = 0.2, 0.1$, $z_{1-\beta} = 0.84, 1.28$), the required sample size is approximately $N = \left[(z_{1-\alpha}/2 + z_{1-\beta}\sqrt{\theta(1-\theta)})/(\theta - 1/2)\right]^2$ (Elandt-Johnson 1971, expression [13.69]). For $z_\alpha = 3$, $\beta = 0.80$, $\theta = 0.05$, one finds $N = 20$, with N being the total number of recombinants and nonrecombinants (Ott 1991, Table 5.8). A rough guide is as follows (Ott 1991, pp257-8): a good picture of the properties of the disease including its phenotype, mode of inheritance, population frequency is obtained first. For a quantitative phenotype, an analysis of mixture of distributions may be required. The number of families (e.g. N=20) required to find linkage to a hypothetical marker in phase-known meioses needs one offspring to each sibship to be added for phase-unknown. To adjust for incomplete marker informativeness, the number of families required is multiplied by the inverse of the marker's heterozygosity. With incomplete penetrance, N must be multiplied by the inverse of the relative expected lod score (ELOD). Heterogeneity may be allowed for by further multiplying the number of offspring by the inverse of the relative efficiency (ratio of variances). When 20% of families are of the unlinked type ($\alpha = 0.80$), with $\theta = 0.05$, the number of offspring must be multiplied by the inverse of the ratio of the variances, that is, by $(0.075/0.042)^2 = 3.2$ (0.075 and 0.042 being the standard errors of θ for α=1 and 0.8 when there are 10 families with 3 phase-known meioses each. Ott 1991, Table 9.7). Errors in marker typing may be allowed for under a simple misclassification model. For

example, with $\theta = 0.05$ and a misclassification rate of 0.05, the number of offspring has to be increased by a factor of 1/0.77=1.3 (0.77 being the relative ELOD of the recombination fraction estimate in the presence of missclassification. Ott 1991, Table 10.3). Therefore $N = 20$ meioses are increased to $20 \times 3 \times 3.2 \times 1.3 \approx 250$ meioses by a penetrance of 50%, a proportion of linked families of 80%, and a misclassification rate of 5%. Power calculation of a proposed linkage study can be formally done using **SIMLINK** (Ploughman & Boehnke 1989), **SLINK** (Ott 1989; Weeks et al. 1990) and more recently **Allegro** (Gudbjartsson et al. 2000), more details and practical examples will be given in Chapter 5.

Lod score analysis is most powerful when parameters of the genetic model are known. Misspecification of these parameters not only decreases power to detect linkage but also introduces bias in recombination estimate, especially when wrongly specified the dominance (Clerget-Darpoux et al. 1986). A traditionally model-free approach is the allele-sharing method, in which excess in allele sharing between affected members in a pedigree is used to test for linkage between a marker locus and a disease locus, in the hope that aggregation in such families tend to be more genetic and less heterogeneous. The simplest allele-sharing method is affected sib-pair method originated from Penrose (1935). Genes shared by relatives from common ancestor are called **IBD**, or identity by descent. The IBD probabilities for sib pairs were derived by Suarez et al. (1978) and a more comprehensive derivation was given by Guo (2000). Allele-sharing is most successfully applied when the marker locus is extremely polymorphic. The so-called affected pedigree member (APM) method (Lange & Weeks 1988) examines allele-sharing in extended pedigree based on identity-by-state (IBS) relationships, $Z_{ij} = 1/4 \sum_{a=1}^{2} \sum_{b=1}^{a} \delta(A_a, B_b) f(A_a)$ where $\delta(.,.)$ is the Kronecker delta function with $\delta(s, s')$ taking values 1 or 0 according to whether or not s and s' shares alleles IBS; $f(A_a)$ is a weight function taking forms of $1, 1/\sqrt{p_{A_a}}$, $1/\sqrt{p_{A_a}}$, the latter two give more weight to rare-allele sharing. Several limitations of APM method have been raised (Shih & Whittemore 2001). First, it is sensitive to misspecification of marker allele frequencies. Second, the null distribution of the test statistic may be skewed, leading to a potential anti-conservative test. Third, it ignores IBD information available and is thus less powerful than methods based on IBD sharing. A nonparametric linkage statistic (NPL) is based on IBD configurations between pairs of relatives (NPLpair) or all relatives (NPLall) (Whittemore & Halpern 1994a, 1994b; Kruglyak et al. 1996). Davis et al. (1997) proposed to use IBD information and implemented their method in computer program **SimIBD**. A summary of

power of allele-sharing method was given by Shih & Whittemore (2001).

Variance component analysis is the study of continuous traits and their variation due to genetic and environmental factors (Hopper & Mathews 1982; Goldgar 1990; Goldgar & Oniki 1992; Schork 1993; Fulker et al. 1995; Beaty 1997; Lange 1997; Almasy & Blangero 1998; Williams et al. 1999; Pratt et al. 2000). Under this model, the variance of a quantitative trait is partitioned into various genetic and environmental components, which may include additive, dominance, shared and nonshared environment. The genetic component may be further partitioned into a QTL (quantitative trait locus) component and residual genetic compoent, when genetic marker data is available. Denote the kth allele at the QTL locus to be a_k, individuals with genotype a_k/a_l has average trait value $\mu_{kl} = \mu_{lk}$, Statistical inference is based on the covariance structure of the quantitative traits between individuals, which is expressed in terms of several components and their correlations. The correlation of additive genetic component is determined by **kinship coefficient** (ϕ_{ij}, the probability of randomly drawing an allele in individual j being IBD to an allele at the same locus randomly drawn from individual i). The correlation of dominance is Δ_{7ij} (Jacquard 1974, the probability that both alleles at the locus being IBD). The correlation of environmental factor is γ_{ij} (the probability of sharing a particular environmental factor). Values for ϕ_{ij} and Δ_{ij} for common pairs of relatives are known, e.g., 1/4 and 1/8, for sibs, and $1/2^{(k+1)}$ and 0 for most other outbred relationships, where k is the degree of relationship. The covariance between relatives i and j with traits X_i and X_j can be expressed as $Cov(X_i, X_j) = 2\phi_{ij}\sigma_a^2 + \Delta_{7ij}\sigma_d^2 + \gamma_{ij}\sigma_c^2 + \delta_{ij}\sigma_e^2$, where σ_a^2 and σ_d^2 are the additive and dominance effects of the determining locus, and δ_{ij} is the Kronecker symbol, i.e. 1 if $i = j$ and 0 otherwise. The $N \times N$ expected variance-covariance matrix for an entire family of N individuals with data $X = (X_1, ..., X_N)$ can be written as $\Omega = 2\Phi\sigma_a^2 + \Delta\sigma_d^2 + \Gamma\sigma_c^2 + I\sigma_e^2$, where $\Phi = \{\phi_{ij}\}$, $\Delta = \{\delta_{ij}\}$, $\Gamma = \{\gamma_{ij}\}$ are written in matrix forms. Assuming X follows multivariate normality, the likelihood function for the pedigree is $(2\pi)^{-\frac{N}{2}}|\Omega|^{-1/2}\exp[(X - E(X))'\Omega^{-1}(X - E(X))]$, where $E(X)$ may be a constant or a linear function of covariates. The total likelihood is the product of all the likelihoods of individual families. Likelihood ratio test can be applied to nested hypotheses about these parameters. Genetic markers can also be incorporated into the model to infer the quantitative trait locus (QTL), say q. Regardless of the dominance and environmental effects, the variance-covariance matrix Ω is given by $\Omega = \hat{\Pi}\sigma_q^2 + 2\Phi\sigma_a^2 + I\sigma_e^2$, where $\hat{\Pi}$ is a matrix whose

elements $\hat{\pi}_{ij}$ are the estimated proportion of genes shared IBD at the QTL by individuals i and j, σ_q^2 being the genetic variance due to QTL, Φ being the kinship matrix for the pedigree, σ_a being the residual additive genetic variance, I being identiy matrix, and σ_e^2 being the variance due to individual-specific random environmental effects. For the model with n QTLs the first term of the right hind side is the sum of similar terms from the n QTLs (Almasy & Blangero 1998). Method of QTL linkage analysis in sib pairs was discussed by by Haseman & Elston (1972), Wright (1997), Elston et al. (2000), Sham & Purcell (2001). A GEE approach was proposed by Amos (1994, 1996).

Power of variance component analysis was investigated by Fulker & Cherny (1996), Wijsman & Amos (1997), Williams & Blangero (1999), Page et al. (1999), Amos et al. (2001). Exact expression for sib pairs was given by Williams & Blangero (1999), in which a test of $\sigma_q^2 = 0$ has noncentrality parameter $\hat{\delta} = q^2(h^4+4)/2/(h^4-4)^2$, where $q^2 = \sigma_q^2/\sigma^2$, $h^2 = (\sigma_q^2 + \sigma_a^2)/\sigma^2$, σ^2 being the total phenotypic variance. For a test of linkage having 80% power and a size of 0.001 (a lod score of 3), the critical value of the χ^2 test statistic is 20.78, so that the number of sib pairs required is $N = 20.78/\hat{\delta}$, i.e., $2N$ individuals. One of the major concerns of both the variance components model and the Haseman-Elston regression methods is the robustness. For example, the result of linkage analysis could depend on families with extreme scores (Amos & de Andrade 2001, Dr Tao Li personal communication).

Computer implementation of linkage analysis has drawn considerable attetion. Following the work of Haldane & Smith (1947), Smith (1953) and Morton (1955), Elston and Stewart (1971) formulated an algorithm for general pedigrees without loops. This was subsequently generalised by Ott (1974) and implemented in **LIPED**, Lange & Boehnke (1975), Lange et al. (1988) in **MENDEL/FISHER**, Lathrop et al. (1984) in **LINKAGE**, and (Cottingham et al. 1993) in **FASTLINK**. Both **LINKAGE** and **FASTLINK** have modules **ILINK**, **MLINK** and **LINKMAP**. Lander-Green algorithm was implemented in **MAPMAKER** and associates **GENEHUNTER** (Kruglyak et al. 1996), **GENEHUNTER PLUS** (Kong & Cox 1997), **Allegro** (Gudbjartsson et al. 2000), **GENEHUNGER/IMPRINTING** and **GENEHUNTER/TWOLOCUS** (Strauch et al. 2000). Likelihood based methods for affected sib pairs incorporating the possible triangle constraint are implemented in **SPLINK** (Holmans 1993). **ESPA** (Sandkuijl 1989) uses **MLINK** to infer incomplete information and does χ^2 test for increased allele sharing

among affected sib pairs. **APM** implements affected pedigree member method (Weeks & Lange 1988). Variance component linkage analysis has been implemented in **MIM** (Goldgar 1990), **SOLAR** (Almasy & Blangero 1998), **ACT** (Amos 1994). Lange (1997) described method for two quantitative traits. Haseman-Elston method has been implemented in **SAGE**, which also includes Bonney's regressive model and IBD calculation of a general pedigree using an extended Lander-Green algorithm.

Owing to the limited number of meioses the resolution of linkage from family data is usually no finer than 1cM (Boehnke 1994). Therefore there is a need (1) to redefine disease and increase the relative risk among relatives. (2) to fine-structure LD mapping using allelic or haplotype association (Kruglyak & Lander 1995; Nevanlinna et al. 1980; de la Chapelle 1993, 1998).

1.5 Association analysis

Association analysis is the search for genetic markers that occur at a different frequency between cases and controls. An allele is said to be positively associated with the disease if it is more frequent in affected probands than in unaffected controls. This could occur for many reasons (Lander & Schork 1994), for example, (1) the allele itself is a cause of the diseases; (2) the allele is in linkage disequilibrium (LD) with a disease causing gene; and (3) the association is an artifact due to population admixture. The simplest test of allele frequency differences between cases and controls is χ^2 test (Sham & Curtis 1995b). If more than one locus is involved this could be extended to haplotype analysis (Xie & Ott 1993; Terwilliger & Ott 1994; Zhao, Curtis & sham 2000). Considerable attention has been focused on case-control design (Morton & Collins 1998; Pritchard & Rosenberg 1999; Pritchard et al. 2000; Long & Langley 1999; Devlin & Roeder 1999; Risch 2000). Successful examples of case-control studies include the human leukocyte antigen (HLA) associations and a number of diseases including insulin-dependent diabetes melitus (IDDM), multiple sclerosis, rheumatoid arthritis, psoriasis and celiac disease. Risch (2000) pointed out the possibility that the high false positive rate from case-control designs may be due to a low prior probability of gene polymorphisms causally related to the disease outcomes.

Population admixture or substructure has been the major concern of association study, as allele or haplotype frequency differences between populations may lead to spurious asso-

ciation (e.g. Mckeigue 1997; Pritchard et al. 2000a, 2000b; Pritchard & Przeworski 2001). For nuclear families with a single affected child, Rubinstein et al. (1981), Falk & Rubinstein (1987) proposed forming controls with parental marker alleles that have not been transmitted to the child. Thomson (1988) proposed a similar method called "affected family based controls (AFBAC)" (implemented in computer program AFBAC). Further investigations were made by Ott (1989), Terwilliger & Ott (1992), Spielman et al. (1993), Knapp et al. (1993). The popular transmission/disequilibrium test (TDT) was coded by Spielman et al. (1993). The TDT method can be applied to large pedigrees with many affected subjects or to a simplex family with an affected subject and his/her parents. However a simplex family is informative for linkage only when linkage disequilibrium exists, i.e., when the likelihoods of the coupling and repulsion linkage phases are unequal. Therefore TDT can be considered as a test for linkage in the presence of linkage disequilibrium. Methods of Schaid & Sommer (1993, 1994) allow one to estimate separately the magnitude of the relative risks of disease for individuals being homozygous and those being heterozygous at the candidate-gene locus. The first is a HWE likelihood method, and the second is a conditional (on parental genotype, CPG) likelihood method, appropriate when HWE is absent. Knapp et al. (1995) showed HWE method to be asymptotically at least as efficient as the CPG method, irrespective of the mode of inheritance. They also obtained the analytical MLEs for the parameters of HWE method and showed the MLEs of CPG method to be the solution of a simple cubic equation. The original TDT has been extended to multiallelic marker by Bickeböller (1995), Sham & Curtis (1995a), Spielman & Ewens (1996), Morris et al. (1997a, b), Wilson (1997), Martin et al. (1997), Lazzeroni & Lange (1998). Schaid (1996) and Sham (1997) both proposed score statistics from logistic regression. Koeleman et al. (2000) proposed variation based on Sham & Curtis (1995a). When information from one parent is unavailable the use of parent-offspring pair may induce bias (Curtis & Sham 1995b; Sun et al. 1999). In the absense of parental information, inference can be based on extra siblings (Curtis 1997; Spielman & Ewens 1998; Boehnke & Langefeld 1998; Horvath & Laird 1998; Schaid & Rowland 1998; Whittaker & Lewis 1999; Siegmund et al. 2000), which is a "horizontal" approach in contrast to "vertical" approach by reconstruction from missing parents (Clayton 1999) or both (Knapp 1999b). Reconstruction of parental information has been extensively discussed (Chiano & Clayton 1998; Weinberg et al. 1998; Weinberg 1999a,b; Umbach et al. 2000). Other developments include the Monte Carlo sim-

ulation (Cleves et al. 1997; Zhao, Sham & Curtis 1999; Whittaker & Thompson 2000), the handling of arbitrary family structures or missing data patterns (Martin et al. 2000; Monks & Kaplan 2000), Whittemore & Tu 2000; Rabinowitz & Laird 2000), haplotype methods (Clayton & Jones 1999; Clayton 1999; Xiong & Jin 2000; Zhao et al. 2000; Dudbridge et al. 2000), and quantitative traits (Allison 1997; Rabinowitz 1997; Fulker et al. 1999; George et al. 1999; Allison et al. 1999; Van den Oord 2000; Waldman 1999; Abecasis, Cardon & Cookson 2000; Rabinowitz & Laird 2000; Laird et al. 2000; Lake et al. 2000; Monks & Kaplan 2000; Zhu & Elston 2001; Zhu et al. 2001). Lunetta et al. (2000) considered general association models for an arbitrary phenotype and score statistics.

An elaboration of Clayton (1999) as in the the computer program **TRANSMIT** is given here. It is "*for estimating genetic associations from probabilities of haplotype transmission to affected offspring when there may be uncertain marker haplotype assignment*". Based on CPG likelihood, the (log)haplotype relative risk parameters with respect to multiplicative model for genotype risk, a score test $U^T V^{-1} U$ is constructed. For known parental haplotypes, U is the difference between the vector of observed marker haplotype frequencies in affected offspring and the frequencies expected under Mendelian transmission, and V is the variance of U (assuming Mendelian transmission). When transmission is fully observed, this test reduces to the usual Pearson χ^2 test. When parental haplotype is uncertain, the observed transmission frequency is averaged over all the possible assignments consistent with the observed data, weighting by their posterior probabilities. The population haplotype frequencies are estimated from the data by EM, while the score test is justified using profile likelihood argument. The information matrix involving all parameters is $I = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, then $V = A - BC^{-1}B^T$ where A and C correspond to haplotype relative risks and haplotype frequencies, respectively. A robust "information sandwich" estimator for the score avoids the need to assume independence of the information from two affected siblings, an assumption viable with large sample size and asymptotic behaviour is unreliable with rare haplotypes, which may be user-controlled. The estimator is derived by assuming the variance of U at the correct model to be $J = \begin{pmatrix} D & E \\ E^T & F \end{pmatrix}$, the variance for U is then obtained from estimating equations as $V = D + BC^{-1}FC^{-1}B^T - BC^{-1}E^T - EC^{-1}B^T$. For the correct model, $J = I$. How common a haplotype must be in order to use the χ^2 tests is argued as follows. Since difference between observed and expected transmission $O - E$

$\sim (n/2, n/4)$, in order for $n/2 > 5$, then $n = 10$ or $n/4 = 2.5$ is needed. For rare haplotypes it offers aggregation but no simulation options. "*The program works by imputation of the missing information concerning parental genotypes and haplotype phase. Inevitably, this brings in a population model and it has been necessary to assume HWE model and no population-admixture...*"

The power to detect association using a marker depends on many factors such as the strength of the linkage disequilibrium between the marker and disease, the age and frequency of the disease mutation, the recombination fraction between the disease and marker, the increase in risk attributatble to the particular disease-susceptibility locus under consideration, and the penetrances of the different disease-locus genotypes. (Terwilliger & Ott 1992; Schaid & Sommer 1993; Weeks & Lathrop 1995; Ott & Rabinowitz 1997; Chapman & Wijsman 1998; Sham, Zhao & Curtis 2000).

Many computer programs have been developed for association analysis. **ASPEX** (Hinds & Risch 1996) contains several programs to perform sib-pair analysis and TDT in nuclear families. For TDT it provides adjustment for multiple testing, while **GENEHUNTER** also provides TDT tests for up to 5 loci. **ETDT** implements Bradley-Terry logistic model for multiallelic marker, and a module **MCETDT** (Zhao, Sham & Curtis 1999) for Monte Carlo simulation. Another associate **LAMBDAA** is written for TDTmax test (Morris et al. 1997). **SIBASSOC** is a derivation of **ETDT** (Curtis 1997). **SDT** implements Spielman-Ewens (1997). **GASSOC** implements multiallelic TDT of Schaid (1996). **XDT** and **FBAT** implements methods by Horvath & Laird (1998), Rabinowitz & Laird (2000), Lake et al. (2000), and Laird et al. (2000). **PDT** implements pedigree disequilibrium test of Martin et al. (2000). **QTDT** implements several quantitative TDTs (Abecasis et al. 2000). A SAS program is available for obtaining Spielman-Ewens, Bickeböller, Stuart statistics (Sham 1997) and for evaluating power of **ETDT**. Combination of TDT and the measured haplotype method has recently been proposed by Seltman et al. (2001). **TRANSMIT** (Clayton 1999) and **MULTDT** (Zhao et al. 2000) were developed for transmission/disequilibrium tests involving multiple marker loci.

The high resolution of association analysis relies on abundance of genetic markers, especially with the availability of SNP (Wang et al. 1998; Lindblad-Toh et al. 2000), which has advantages and disadvantages over traditional markers (Kwok & Gu 1999): First, the mean density of SNPs is approximately one per kilobase in the human genome, so that a SNP

library will be sufficient for various study designs. Second, the mutation rate per generation of SNPs is low, which makes it the best choice for association studies. Third, it is likely that a subset of SNPs are functionally important in etiology of complex traits, given that many Mendelian disorders result from single nucleotide change. Fourth, low-cost, high-throughput, automatic genotyping methods are available for efficient genotyping, making it possible to conduct large population studies. In contrast, microsatellite markers are rare, with higher mutation rate and less amenable to high-throughput genotyping. However, methods that could handle massive data sets are not yet well-developed. Without a priori knowledge of genotyping errors the conclusion may be biased. SNP haplotypes will have to be constructed to match the information content of microsatellites in association studies. As illustrated by Nickerson et al. (1998) and Clark et al. (1998) in their study of lipoprotein lipase (LPL) gene, 9.7kb of genomic DNA within this gene reveals 88 variable sites among 71 healthy individuals. Most of the variability occurs in noncoding regions, with a high rate of interlocus recombination and a complex pattern of linkage disequilibrium. It was not possible to assume any SNP would '*give reliable information about flanking sites*' and that a random sampling of three of four SNPs in the 10kb analysed 'would not be a reliable method of detection of nearby causal variation'. If these effects are typical, they will result in considerable loss of power to detect any predisposing disease alleles. The characterisation of sequence variation across genomic regions in general is far from over (Subrahmanyan et al. 2001).

1.6 Coalescent models

The completion of DNA sequencing should help to understand the biological mechansims underlying human traits and offer opportunities to decipher enigmas of evolution. Coalescent theory (Felsenstein 1981; Kingman 1982, 2000; Hudson 1990; Lange 1997; Hartl & Clark 1997) in particular is recognised as a cornerstone for various statistical analyses of molecular population data (Fu & Li 1999): First, it is a sample-based theory, proposing that description of a sample is more relevant than that of a whole population. Second, it is efficient and motivates many algorithms for simulating population samples under various population genetics models. Third, it is particularly suitable for molecular data, such as DNA sequence samples.

Consider a sample of n sequences of DNA region from a population, and assume there is no recombination between sequences. These sequences are connected by a single phylogenetic tree or genealogy, the root of the tree is the most recent common ancestor (MRCA). Under the assumption of neutrality and when n is small relative to the total population size N, Kingman (1982) showed that, to a first approximation, only two lineages coalesce at each event and that the distribution of times between successive coalescent events has a geometric distribution $P(t|i) = [1-i(i-1)/4N]^{t-1}[i(i-1)/4N] \approx \exp[-i(i-1)t/4N][i(i-1)/4N]$, where i is the number of lineages remaining and t is the number of generations until the next coalescent event. The probability of any topology of the gene genealogy can be generated from the assumption that when i lineages remain, any of the $i(i-1)/2$ possible coalescent events is equally likely to occur. This process for two sequences of diploid organisms under the neutral Wright-Fisher model can be described as follows (Fu & Li 1999; Slatkin 1999; Hartl 2000). Looking backward in time, the probability of two alleles in the present generation having distinct ancestors in the previous generation is $1-1/(2N)$. The probability of a third allele having an ancestor distinct from the first two is $1-2/(2N)$. The overall probability that i distinct alleles present in any generation have i distinct ancestors in the previous generation is $[1-1/(2N)][1-2/(2N)]\ldots[1-(i-1)/(2N)] \approx 1-(1+2+\ldots+i-1)/(2N) = 1-i(i-1)/(4N)$, so that the probability of the presence of a coalescence is $i(i-1)/(4N)$. Therefore for i alleles, the probability of no coalescence for the first $t-1$ generations followed by coalescence in the tth generation is $[i(i-1)/(4N)](1-[i(i-1)/(4N)])^{t-1}$, and the mean of this geometric distribution is $4N/[i(i-1)]$. In a given time period the number of mutations that have occurred on a sequence is a Poisson variable, implying that the mean number of mutation in the two sequences is $\theta = 4N\mu$, where μ is the mutation rate per sequence per generation. Now consider the effect of mutation by sample size of $n = 2$, for which the distribution of times until they coalesce is a geometric distribution with mean $2N$ and $i = 2$ given above. For two alleles to be the same state, either there is no mutation before the coalescence or mutations have occurred in such a way that the two copies are in the same site. For an infinite alleles model in which every mutation destroys genetic identity and no mutation can recover it, the probability of no mutation occurring before coalescence at t is $\exp(-2\mu t)$, where the 2 is due to two lineages on which mutation can occur. So the probability of identity is obtained by averaging over the distribution of t, $P = (1/2N)\int \exp(-t/2N - 2\mu t)dt = 1/(1+4N\mu)$ (Hudson 1990).

Alternatively assuming the stepwise mutation model (Ohta & Kimura 1973), two copies of the locus will be in the same state if the mutations on the two branches joining them yield no net change in allele size. Let j_+ and j_- be the number of mutations increase or decrease allele size by one, both with Poisson(μt) distribution because the total length of the genealogy is $2t$, but only half of the mutations will increase or decrease allele size. Given t, the probability that $j_+ = j_-$ is the probability of two draws from Poisson(μt), $P(j_+ = j_-|t) = \sum_{j=0}^{\infty}[(\mu t)^j/j! \exp(-\mu t)]^2 = I_0(2\sqrt{\mu t})$, where $I_0(.)$ is a modified Bessel function, thus the total probability of identity is obtained by averaging over t, $P = 1/\sqrt{1+8\mu t}$. Variations in population size will change the distribution of coalescent times, but can be absorbed into a transformation of time scale in the coalescent process for a population of constant size.

The original form by Kingman (1982) for a neutral locus in a large, random mating population with constant population size has been well extended for variable population size, population structure including island model, stepping stone model, island-continent model, selection (Neuhauser & Krone 1997; Krone & Neuhauser 1997), recombination (Griffiths & Marjoram 1996), and conditional coalescent. The inference concerns the topology of the evolutionary tree, as effect of different evolutionary forces such as mutation, recombination, selection and migration (Tavaré et al. 1997; Wilson & Balding 1998) as in their programs **micsat** and **BATWIN**. Computational methods involve recursive equations or importance sampling (Griffiths & Tavaré 1994a,b; Felsenstein et al. 1999), MCMC (Kuhner et al. 1995, 1998), among others (Bahlo & Griffiths 2000; Nielsen 1997, 2000; Wiuf 2000). Stephens & Donnelly (2000) suggested an improved importance sampling scheme for histories of mutation and coalescence events. These methods have been implemented in **genetree** (ftp://ftp.monash.edu.au), and **lamarc** (Likelihood Analysis with Metropolis Algorithm using Random Coalescence: **coalesce**, **fluctuate**, **migrate**, **recombine**, http://evolution.genetics.washington.edu/). Many applications of coalescent theory have been to samples from the control region of human mtDNA. Since mtDNA is haploid with recombination, DNA polymorphism is available naturally in haplotypes. Similar application is to non-recombining region of human Y chromosome (Cooper et al. 1996, Wilson & Balding 1998). The study on the β-globin gene tree by Harding et al. (1997) represents a notable example of human population study using a nuclear locus. Again due to high mutation rate of microsatellite marker it is less ideal for evolutionary studies. The relative

importance of various evolutionary forces such as recombination, natural selection, migration and population growth are not well understood. Molecular evolution, protein and DNA sequence analysis are now within the framework of bioinformatics. See Searls (2000) for an overview, Pietro & Goldman (1998), Fortna & Gardiner (2001), Whelan et al. (2001) for discussions of computer programs.

1.7 Outline of thesis

This thesis will focus on several scenarios of statistical power analyses. For linkage analysis this includes several test statistics and the use of homozygosity mapping via computer simulation. For association analysis this includes model-free statistics for case-control design and tests of marker polymorphism and mutation detection. A closely related investigation is the combined linkage and association under likelihood framework. Power analysis of both linkage using affected sib-pairs and association using TDT as in Risch and Merikangas (1996) is also given. Chapter 2 presents power comparison of parametric and nonparametric linkage test statistics for small families. Chapter 3 investigates power of model-free statistics in case-control association design. Chapter 4 studies effect of marker characteristics on mutation detection. Chapter 5 describes a simulation program for oligogenic traits, which is compared with simulation methods with conditioning and homozygosity mapping analyses. Chapter 6 collects some numerical experiments on haplotype analysis in nuclear families. Chapter 7 examines power of TDT as compared to other designs. Chapter 8 is a discussion of the main results, remaining work and some general issues. Most chapters consist of introduction, methods, results, discussion, and bibliographic notes. The basic of gene counting method and computer programs used in relevant chapters are given in the Appendix.

1.8 Bibliographic notes

The major reference for the background materials is Thompson (1996). Some definitions of statistical genetics were also given by Elston (2000). A historical review of genetics was given by Crow (2000). Other references included Elandt-Johnson (1971), Morton et al. (1983), Khoury et al. (1993), Weiss (1993), Pawlawitzki et al. (1994), Lander & Waterman (1995), Thompson (1995), Speed & Waterman (1996), Weir (1996), Hartl & Clark (1997), Lange (1997), Gillespie (1998), Sham (1998), Ott (1999) and Hartl (2000).

Other reviews include Lander & Schork (1994), Thompson (1995), Weeks & Lathrop (1995), Schork et al. (1998), Olson et al. (1999), Thomson & Esposito (1999), Morton (2000), Risch (2000), Guo & Lange (2000). Methods for experimental genetics have been discussed in Doerge et al. (1997), Darvash (1998), Moore & Nagle (2000). General strategy has been discussed in Zhao et al. (1997). Literature concerning other commonly used study designs, such as migration, twin and adoption studies is not elaborated here. A comprehensive bibliography and list of computer programs for linkage and association anlaysis are available via http://linkage.rockefeller.edu.

The thesis is a summary of several papers (e.g. Sham, Lin et al. 2000, Chapter 2; Zhao & Sham 1997, Zhao, Curtis & Sham 2000, Chapter 3; Sham, Zhao & Curtis 2000, Chapter 4; Ohadi et al. 1999, Chapter 5) and personal communications (Dr Anthony D. Long, Chapter 7; Dr Mark Layton, Dr Duncan Thomas and GAW12 RFA, Chapter 5; Dr Momiao Xiong, Chapter 6; Dr Carlos Zapata, Appendix A). Notes elaborating MCMC linkage analysis, affected sib pair (Sham & Zhao 1998), QTL analysis, and a number of computer programs are available from Section of Genetic Epidemiology and Biostatistics via http://web1.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/index.shtml and my personal web pages http://www.hgmp.mrc.ac.uk/~jzhao and http://www.ucl.ac.uk/~rmjdjhz.

A brief description of work other than my own is given as follows.

- Chapter 2. This is a continuing work of Lin et al. (1997) at a referee's suggestion of looking at a fully informative marker. The formal definitions and distributions of statistics MLOD, MFLOD, MALOD are due to Professor Sham and Dr Curtis. Dr Curtis also helped to clarify 1-sided versus 2-sided of these tests.

- Chapter 3. The use of Equation (3.1) for simulation is due to professor Sham. The power calculation is a joint work with Professor Sham.

- Chapter 4. Equation (4.1) is due to Professor Sham.

- Chapter 5. Program **SIM** is a joint work with Professor Sham.

- Chapter 6. The possibility to use haplotypes in **LINKAGE** to explore LD was suggested by Dr Curtis. Professor Sham also suggested Equation (3.1) for model selection and reparametrisation of the optimisation procedure.

Chapter 2

Linkage tests for small pedigrees

2.1 Introduction

Linkage analysis of complex traits is commonly performed using parametric lod score method and nonparametric allele-sharing method. Parametric tests generalise the traditional lod score method by maximising over recombination parameters while treating allele frequencies and mode of inheritance as nuisance parameters (Risch 1984; Clerget-Darpoux et al. 1986; Elston 1994; Hodge & Elston 1994). The MOD score is defined as the maximum lod score over mode of inheritance. Minor misspecification of mode of inheritance can affect estimation of recombination parameters but not seriously reduce lod score (Clerget-Darpoux et al. 1986). However, it has the undesirable property that its distribution is unknown under the null hypothesis, with penetrance playing a role in the alternative hypothesis but not defined under the null. Curtis & Sham (1995a) attempted to amend this by introducing MFLOD (model-free lod score), obtained via the computer program MFLINK, where maximised lod score and maximised admixture lod score were also calculated. Interestingly both were observed approximately proportional to χ^2 statistics and occasionally larger than MFLOD. Lin et al. (1997) therefore named them as MLOD (lod score maximised over disease models) and MALOD (admixture lod score maximised over disease models). Nonparametric allele-sharing statistics are mostly generalisation of sib-pair statistics (Suarez et al. 1978, Blackwelder & Elston 1985) by using IBS (Weeks & Lange 1988) and IBD information among all affected relative pairs (Whittemore & Halpern 1994a, 1994b; Kruglyak et al. 1996).

The performance of these statistics depends on family structure, number of affected in-

dividuals, mode of inheritance, location and informativeness of markers (e.g., Boehnke 1991; Ott 1992; Terwilliger & Ott 1994). In addition to mode of inheritance and recombination rate between markers and the disease genes, power of allele-sharing methods also depends on how allele-sharing within a pedigree is scored and how these scores are weighted across pedigrees. These have generated considerable interests (Durner et al. 1992; Risch & Zhang 1995; Davis et al. 1996; Sham et al. 1997; Davis & Weeks 1997; Hodge 1997; Hodge 1998; Goldin, Greenberg et al. 1998; Durner et al. 1999; McPeek 1999; Blackwelder & Elston 1985; Risch 1990a,b,c; Holmans 1993; Kruglyak & Lander 1995; Hauser et al. 1996; Whittemore 1996; Kruglyak et al. 1996; Kong & Cox 1997; Hauser & Boehnke 1998; McCarthy et al. 1998; Abreu et al. 1999; Teng & Siegmund 1998; Dudoit & Speed 2000; Dupuis & van Eerdewegh 2000; Liang et al. 2001; Shih & Whittemore 2001). Unfortunately, Sengul et al. (2001) noted that it is hard to make direct comparison of these investigations, owing to differences in these study designs.

Lin et al. (1997) used computer simulation and simple family types to evaluate the performance of parametric and NPL statistics under single gene models with linkage homogeneity and heterogeneity and concluded that, under a model with a major gene effect, likelihood-based methods tend to be more powerful. For a minor gene effect, the NPL statistics were generally superior to the other tests.

Chapter aims

This chapter concerns the comparison of power of parametric and nonparametric linkage tests under complete marker information. The parametric and non-parametric linkage test statistics include MLOD, MALOD, MFLOD, and NPLs. Four types of simple pedigrees and their mixture and several modes of inheritance are used to examine behaviour of these statistics under both linkage homogeneity and heterogeneity. In contrast to Lin et al. (1997), the power comparisons will be based on exact calculations.

2.2 Methods

Definition of test statistics

We restrict ourselves to the usual generalised single locus model, in which the disease locus is biallelic consisting of both a disease allele and a normal allele. Note there are other single

locus models, for instance those with multiple alleles (Nielsen, Ehm & Weir 1998; Nielsen & Weir 1999, 2001). We also allow for mixture of families exist in the study sample, in which linkage only occurs in a subset of families. We study several two-point linkage statistics under these assumptions.

Let q = the frequency of the disease allele; f_i = the penetrances, namely the conditional probabilities of the disease given i copies of the disease allele, $i = 0, 1, 2$; θ = the recombination fraction between the disease locus and the marker locus; α = the proportion of families with linkage. The lod score and admixture lod score under this disease model are denoted as **LOD** and **ALOD**, respectively. To calculate MLOD, MALOD and MFLOD, disease model parameters are constrained to produce the correct population disease prevalence (K),

$$\begin{aligned} f_0 = f_1, f_2 = 1 - f_1(1-K)/K \quad \text{when} \quad f_1 \leq K \\ f_2 = f_1, f_0 = (1-f_1)K/(1-K) \quad \text{when} \quad f_1 > K \end{aligned} \tag{2.1}$$

which implies that only f_1 is free, and that and q is given by solving $q^2 f_2 + 2q(1-q)f_1 + (1-q)^2 f_0 = K$ (Curtis & Sham, 1995). Specifically, f_1 is restricted to vary along the sides of the triangle of $(0, 0, 1)$, (K, K, K) and $(0, 1, 1)$, so that all the parameters are well characterised by f_1 and K. By using marker data (M) and disease phenotype (D), MLOD, MALOD and MFLOD can be evaluated as a function of f_1 and α at a test position $\theta = t$, as follows (Sham, Lin et al. 2000).

MLOD: the maximum lod score obtained over all feasible transmission models, i.e. for a given value of $\theta = t$,

$$\begin{aligned} & \max_{f_1} \log_{10}(P(D, M; \theta = t, f_1)/P(D, M; \theta = 0.5, f_1)) \\ = \; & \max_{f_1} \log_{10}(P(M|D; \theta = t, f_1)/P(M)) \\ = \; & \max_{f_1} \log_{10}(P(M|D; \theta = t, f_1)/P(M|D; \theta = t; f_1 = K)) \end{aligned}$$

with α being fixed to be 1. For instance, the second step holds since under no linkage (i.e., $\theta = 0.5$) the disease locus and the marker are independent. The only free parameter associated with MLOD is f_1, which is allowed to vary between 0 and 1, with the null hypothesis value being $f_1 = K$. MLOD is therefore the logarithm of ratio of the likelihoods of two nested models, and $(2 \ln 10)$MLOD is asymptotically distributed as χ^2_1 (2-sided, since f_1 can be greater or less than K).

MALOD: the maximum admixture lod score (maximum over these transmission models and α), i.e,

$$\begin{aligned} & \max_{\alpha,f_1} \log_{10}(P(D,M;\theta=t,\alpha,f_1)/P(D,M;\theta=0.5,\text{ or }\alpha=0,f_1)) \\ = & \max_{\alpha,f_1} \log_{10}(P(M|D;\theta=t,\alpha,f_1)/P(M)) \\ = & \max_{\alpha,f_1} \log_{10}(P(M|D;\theta=t;\alpha,f_1)/P(M|D;\theta=t,\alpha=0 \text{ or } f_1=K)) \end{aligned}$$

MALOD statistic is characterised by two free parameters, f_1 and α, which are fixed and completely confounded under the null hypothesis (which can be specified by either $f_1 = K$ or $\alpha = 0$), therefore $(2\ln 10)$MALOD is somewhat conservative for an asymptotic distribution of a $50:50$ mixture of χ^2_0 and χ^2_2, under the null hypothesis that linkage is absent, and that it would be more unfavourable to refer MALOD to χ^2 distribution with two degrees of freedom and would yield a conservative test.

MFLOD: the lod score obtained from the difference between likelihood maximising over f_1, α and likelihood maximising f_1 but setting α to be zero, i.e.,

$$\max_{\alpha,f_1} \log_{10}(P(D,M|\alpha,\theta=t,f_1)/\max_{f_1} P(D,M|\alpha=0,\theta=0,f_1))$$

Asymptotically, $(2\ln 10)$MFLOD is χ^2 with one degree of freedom as the only parameter (free in the numerator likelihood but fixed to 0 in the denominator likelihood) is α, which differs from MLOD and MALOD in that it is the logarithm of a ratio of the joint likelihoods of marker and disease phenotypes, rather than the logarithm of a ratio of the conditional likelihoods of marker phenotypes given disease phenotypes.

NPL statistics: which include NPLpair and NPLall and are based on normal approximations of score functions S_{pair} and S_{all}. Let n_A be the number of affected relatives in a pedigree, h a collection of alleles generated by taking one allele from each affected individuals (there are 2^{n_A} possible collections), $2f$ the total number of founder alleles (total number of IBD alleles) in the pedigree, $b(h)$ the total number of appearances of a founder allele in the collection h, the two score functions assign numerical values to all possile IBD configurations,

$$S_{pair} = 2/[n_A(n_A-1)] \sum_{1\le k\le l\le n_A} \left[\frac{1}{4}\sum_{a=1}^{2}\sum_{b=1}^{2}\delta(s_{ka},s_{lb})\right]$$

and

$$S_{all} \quad = \quad 2^{-n_A} \sum_h \left[\Pi_{i=1}^{2f} b_i(h)! \right]$$

where $\delta(.,.)$ is the Kronecker delta function with $\delta(s,s') = 1$ or 0 according to whether or not alleles s and s' are IBD, and $(s_{11}, s_{12}, \ldots, s_{n1}, s_{n2})$ is the inheritance vector indicating which of the $2f$ founder alleles the nonfounders inherit. With score function S_i thus defined, NPL is then obtained as $Z = \sum_{i=1}^{m} Z_i/\sqrt{m}$, $Z_i = (S_i - \mu_i)/\sigma_i$, with m being number of pedigrees in the sample, μ_i and σ_i being the null mean and null standard deviation of S_i. In words, NPLpair is based on number of pairs of alleles from distinct pedigree members are IBD. NPLall is based on the average number of permutations that preserve a collection h (Kruglyak et al. 1996; Sham 1998; Shih & Whittemore 2001).

Genetic models

Four models of major genes with minor effects are considered (see Table 2.1). For each model, the table gives the penetrance (f_i), prevalence (K), and the conditional probabilities (c_i) of having i copies, $i = 0, 1, 2$, of the disease alleles among affected individuals. Two of the models have high penetrance genes. These are common recessive (CR) and common dominant (CD). The use of common recessive and dominant models conforms to the notion that in linkage analysis of complex traits it is preferable to consider both dominant and recessive models (Clerget-Darpoux et al. 1986; Hodge et al. 1997). The other two models are multiplicative (MM1 and MM2) with genotypic relative risks 4 and 3, respectively. The first model, MM1, has been used in Sham & Curtis (1995a); it also gives the same prevalence as MM2.

Table 2.1: The four genetic models

Model	f_2	f_1	f_0	q	K	c_2	c_1	c_0
Common recessive (CR)	.50	.005	.005	.100	.01	.503	.090	.407
Common dominant (CD)	.50	.500	.005	.005	.01	.001	.501	.498
Multiplicative model 1 (MM1)	.80	.200	.050	.130	.10	.140	.468	.392
Multiplicative model 2 (MM2)	.45	.150	.050	.207	.10	.193	.493	.315

Pedigree configurations and marker information

Four typical family types (Figure 2-1) and their mixture are examined. This allows for factors such as number of affected/unaffected individuals in a family and the availability of distant relatives to be examined. Families 1-3 are nuclear families, each containing an affected sib pair, with an extra affected sib in family 2 and an extra unaffected sib in family 3. Family 4 is extended to have a affected cousin with an unaffected sib, the trait phenotypes of unaffected individuals in this family are assumed to be unknown. The proportions of four families in the family mixture are 0.5, 0.2, 0.2, 0.1, respectively.

To control for effect of allele frequencies, a fully informative marker is assumed. To focus on comparison of test statistics, the recombination rate (θ) is assumed to be 0. Marker phenotypes are assumed to be available to all members of the families. There are 4, 16, 16, and 256 possible marker genotype configurations (n, shortened for genotype configurations below) for family types 1-4.

Method of computing asymptotic distributions

Under no linkage, each genotype configuration occurs with equal probability $(1/n)$, compared to p_i^1 under linkage and $(p_i^1 + 1/n)/2$ under admixture of α=0.5, $i = 1, \ldots, n$. These probabilities are used to calculate the expected values of the log-likelihoods, LOD scores and ALOD scores, over fine grid values of f_1 and α. Appropriate values maximised over these grids provide estimate of noncentrality parameter estimates per pedigree for MLOD (f_1), MALOD (f_1 and α) and MFLOD (difference between log-likelihoods maximised over f_1, α and over f_1). The means and standard deviations of all statistics over all possible genotype configurations are calculated under the null hypothesis of no linkage and under alternatives of linkage and linkage heterogeneity.

For a normally distributed test statistic with mean 0 and variance 1 under the null hypothesis and mean m and variance v under the alternative, the sample size required based on 1-sided normal approximation for type I and type II error rates $\alpha = 0.0001$ and $\beta = 0.1$ is $N = \lceil [(3.719 + 1.282\sqrt{v})/m]^2 \rceil$, where the ceiling function $\lceil . \rceil$ always rounds fractional argument up by 1. This has been used for NPL, so has been for LOD and ALOD, as they are calculated directly without maximisation while Central Limit Theorem would justify its use given large number of families. For MLOD, MALOD and MFLOD the required sample

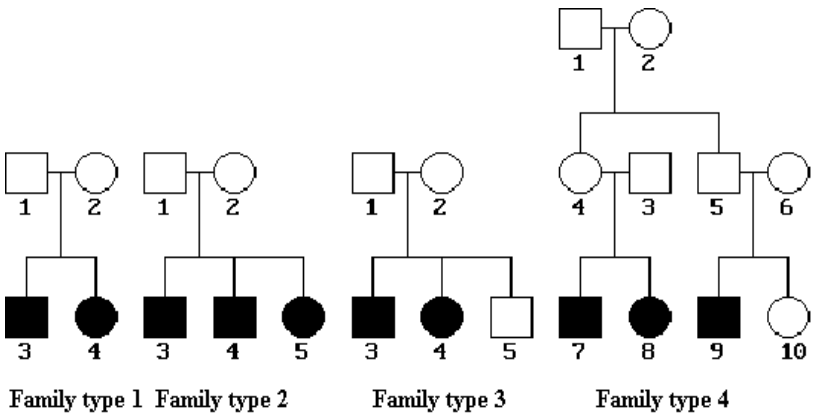

Figure 2-1: The four family types

sizes are calculated according to their noncentrality parameters per pedigree. The required sample sizes are obtained for type I error rates α=0.0001, 0.00001 and 90% power. The noncentrality parameters for χ^2 distribution with degrees of freedom of 1 and 2 are 26.76 and 29.92 for 2-sided tests (25.00 and 28.11 for 1-sided).

The mean and variance of a statistic for family mixture can be obtained from means and variances of individual families over all genotype configurations, as $\sum_{j=1}^{4} \tau_j m_j$ and $\sum_{j=1}^{4} \tau_j (v_j + m_j^2) - (\sum_{j=1}^{4} \tau_j m_j)^2$, where τ is the proportion of family type j in the mixture, and m_j and v_j are the mean and variance of a statistic for family type j, respectively.

Simulation studies

Empirical distributions for these statistics under the null hypothesis were obtained via computer simulation. Replicates are obtained from multinomial distributions of genotype configurations given different family types and transmission models under the null hypothesis. The numbers of replicates are determined by LOD and ALOD such that according to asymptotic theory, a LOD score or a ALOD score test should have 90% power at significance 0.0001. Test statistics are calculated for 10,000 simulated samples in order to provide their empirical distributions. These distributions are compared with the normal or χ^2 distributions at the critical value for a significance level of 0.001, to see if the tests produce correct p values.

We shall illustrate the calculations with family 1, which has four possible genotype configurations. If the parental genotypes are (12, 34) and the first offspring is assigned

genotype 13, then the four configurations are specified by the second offspring (13, 14, 23, 24). We denote these configurations are denoted as $i = 1, \ldots, 4$, so that the following quantities can be obtained,

l_i - likelihood of linkage under a true model

u_i - likelihood of no linkage

p_i^0 - probability of genotype configuration under no linkage

p_i^α - probability of genotype configuration under admixture

p_i^1 - probability of genotype configuration under full linkage

lod_i - lod score for a given genotype configuration

$alod_i$ - admixture lod score for a given genotype configuration

We use **VITESSE** (O'Connell & Weeks 1995; O'Connell 2001) to obtain the quantities $\log_{10}(l_i)$ and $\log_{10}(u_i)$. Then

$$\begin{aligned} p_i^0 &= 1/4 \\ p_i^1 &= l_i / \sum_{j=1}^{4} l_j = 10^{\log_{10}(l_i)+M} / \sum_{j=1}^{4} 10^{\log_{10}(l_j)+M}, \quad i = 1, \ldots, n \\ p_i^\alpha &= (1-\alpha)p_k^0 + \alpha p_k^1 \end{aligned}$$

where M is a large number used to avoid underflow. As usual, $lod_i = \log_{10}(l_i/u_i)$, $alod_i = \log_{10}((\alpha l_i + (1-\alpha)u_i)/u_i) = \log_{10}(\alpha 10^{lod_i} + (1-\alpha))$. Further we have under no linkage

$$\begin{aligned} \text{E(LOD)} &= \sum_{j=1}^{4} p_j^0 lod_j \\ \text{V(LOD)} &= \sum_{j=1}^{4} p_j^0 lod_j^2 - [\text{E(LOD)}]^2 \\ \text{E(ALOD)} &= \sum_{j=1}^{4} p_j^0 alod_j \\ \text{V(ALOD)} &= \sum_{j=1}^{4} p_j^0 alod_j^2 - [\text{E(ALOD)}]^2 \end{aligned}$$

and under linkage

$$\begin{aligned} \text{E(LOD)} &= \sum_{j=1}^{4} p_j^1 lod_i \\ \text{V(LOD)} &= \sum_{j=1}^{4} p_j^1 lod_j^2 - [\text{E(LOD)}]^2 \end{aligned}$$

$$\begin{aligned} \mathrm{E(ALOD)} &= \sum_{j=1}^{4} p_j^{\alpha} alod_j \\ \mathrm{V(ALOD)} &= \sum_{j=1}^{4} p_j^{\alpha} alod_j^2 - [\mathrm{E(ALOD)}]^2 \end{aligned}$$

Using these quantitities the required sample size at significance level of 0.0001 and 90% power can be calculated for LOD and ALOD, which will be used in setting their thresholds in simulations under the null hypothesis and for calculation of MLOD, MALOD and MFLOD. Note that under the null hypothesis means and variances of LOD and ALOD may deviate slightly from 0 and 1, the calculation of sample sizes takes into account of this fact by appropriate standardisation according to their means and variances under the null. Also note that calculations for LOD and ALOD are carried out under true models of linkage and no linkage and we assume $\alpha = 0.5$. The calculations of NPLpair and NPLall are similar to LOD and ALOD as they do not need model assumption, while calculations of MLOD, MALOD, MFLOD allow for variation of f_1 (by keeping prevalence K constant) and α. To compensate for the observation that the maximised log-likelihood for MFLOD often deviates from the true model, a finer grids for f_1 are used either around K, or near 0 and 1 with an exponential degradation, and MLOD and MALOD are obtained accordingly using this grids. MFLOD and MALOD are obtained by modifying **MFLINK**, while NPL statistics are based on **GENEHUNTER** routines *score_pairs* and *score_all*. Indeed the probabilities of genotype configurations for family 1 can be compared to IBD distribution of affected sib pair derived by Suarez et al (1978). The thresholds used in simulation under the null hypothesis for MLOD, MALOD, MFLOD and NPLs are according to χ^2 and normal deviates, respectively.

The whole procedure is implemented in a C computer program. In the implementation, mean $\overline{X}_k$ and variance V_k are iteratively calculated. Start from $\overline{X}_k = X_1$ and $V_k = 0$ for k=1, we have for the kth iteration $\overline{X}_k = [(k-1)\overline{X}_{k-1} + X_k]/k$, $(k-1)V_k - (k-2)V_{k-1} = X_k^2 - k\overline{X}_k^2 + (k-1)\overline{X}_{k-1}^2 \equiv T$, $V_k = [(k-2)V_{k-1} + T]/(k-1)$. Results are obtained under DEC Alpha and Sun SPARC stations.

2.3 Results

Estimated sample sizes required for significance level $\alpha = 0.0001$, power $1 - \beta = 90\%$ at complete linkage ($\theta = 0$) are given in Table 2.2. As expected LOD needs the smallest sample size, although sample sizes required by other statistics are not much larger. The patterns of sample sizes are also as expected. Sample sizes required by multiplicative models MM1 and MM2 may be many folds of models CR and CD. NPL statistics have better performance than MLOD under the multiplicative models considered. Both NPLs and MLOD outperform MALOD and MFLOD.

Table 2.2: Estimated sample sizes required under linkage homogeneity (100% families are linked)

Model	Family	LOD	NPLpair	NPLall	MLOD	MALOD	MFLOD
CR	1	18	20	20	22	24	32
	2	10	10	10	13	15	17
	3	16	19	19	20	22	42
	4	13	14	14	15	17	48
	mixed	15	16	17	18	20	34
CD	1	50	52	52	51	57	51
	2	16	18	18	20	22	20
	3	44	52	52	45	50	72
	4	7	9	7	10	11	13
	mixed	28	32	31	29	32	39
MM1	1	285	285	285	312	348	311
	2	80	80	80	91	102	94
	3	245	312	312	275	307	352
	4	124	131	132	145	162	280
	mixed	173	192	192	195	218	264
MM2	1	626	626	626	677	755	675
	2	199	199	199	224	250	224
	3	582	641	641	636	710	1022
	4	300	309	315	342	382	771
	mixed	405	437	438	449	502	666

("mixed" refers to 50%, 20%, 20% and 10% mixture of family types 1-4)

Table 2.3 gives estimated sample sizes under admixture with 50% of families demonstrating linkage. More families are required compared to full linkage, in a similar pattern

with respective to pedigree types and disease models. The sample size required is more than doubled relative to complete linkage. ALOD will need the smallest sample size over all models and family types. Again NPL statistics perform very closely to but better than MLOD, except for model CR. Overall, all five statistics yield comparable power.

Table 2.3: Estimated sample sizes required under linkage heterogeneity (50% families are linked)

Model	Family	ALOD	NPLpair	NPLall	MLOD	MALOD	MFLOD
CR	1	74	92	92	84	94	119
	2	35	45	45	44	48	63
	3	65	87	87	77	84	172
	4	44	57	57	56	58	157
	mixed	57	73	73	72	74	131
CD	1	227	228	228	241	270	241
	2	79	80	80	88	99	88
	3	191	228	228	224	228	284
	4	26	40	33	35	37	47
	mixed	111	138	133	133	134	149
MM1	1	1139	1151	1151	1245	1390	1243
	2	305	315	315	345	379	357
	3	965	1258	1258	1105	1206	1376
	4	468	512	509	538	599	1008
	mixed	663	764	763	755	841	1041
MM2	1	2503	2515	2515	2701	3018	2703
	2	771	781	781	864	960	860
	3	2317	2577	2577	2514	2806	4036
	4	1159	1212	1224	1288	1439	2887
	mixed	1585	1739	1740	1761	1964	2675

("mixed" refers to 50%, 20%, 20% and 10% mixture of family types 1-4)

Simulation result under the null hypothesis is shown in Table 2.4. Since the significant level has been set to 0.001, we expect 10 of the 10,000 replicates would yield values as extreme under the null hypothesis (above or below this level reveals anticonservative or conservative test). The normal approximations by LOD and ALOD seem acceptable except for model CR. NPL statistics seem to be more anticonservative, especially for family type 4 under models CR and CD, although it is less serious for MM2. For larger pedigrees, NPLpair

seems less liberal than NPLall. Unsatisfactory behaviour under the null hypothesis would make it difficult to compare power, especially for model CR. Although model CR needs smaller sample size than model CD, it is somewhat liberal. MALOD and MFLOD are too conservative overall. This suggests we should choose MLOD over all models.

Table 2.4: Proportion of replicates (out of 10,000) simulated under the null hypothesis reaching threshold of 0.001

Model	Family	LOD	ALOD	NPLpair	NPLall	MLOD	MALOD	MFLOD
CR	1	21	13	7	7	18	3	14
	2	25	26	41	41	10	2	6
	3	20	14	11	11	14	6	7
	4	16	35	21	25	10	3	11
CD	1	4	8	10	10	8	1	7
	2	7	9	11	11	5	0	3
	3	5	10	10	10	10	3	5
	4	37	43	35	57	15	9	4
MM1	1	5	6	5	5	2	0	1
	2	22	14	22	22	8	1	6
	3	8	12	7	7	5	2	5
	4	12	12	12	17	9	1	7
MM2	1	6	15	6	6	6	1	5
	2	9	9	9	9	3	1	3
	3	8	13	12	12	5	5	5
	4	5	20	9	9	6	2	3

2.4 Discussion

The power of the MLOD, MALOD, NPLall and NPLpairs statistics appears to be similar for the types of families considered, which suggests uncertain mode of inheritance is not a serious issue for linkage analysis of genes with minor effects. Compared to findings of Lin et al. (1997), NPLpair and NPLall seem to have better performance but at the same time be more anticonservative. One aspect not revealed by assuming full marker informativeness in this investigation was pointed out elsewhere (Whittemore 1996; Kong & Cox 1997), i.e., NPL statistics could be conservative between markers and more appropriate statistics could be used. In general, to constrain the disease models by the prevalence (K) may be necessary

but not sufficient. For MM1 and MM2 models, the search over f_1 as in (2.1) ignores their multiplicativity and is only a matter of convenience and may tend to underestimate the power of MLOD, MALOD and MFLOD. This highlights the difficulty in practical data analysis, for the disease models are often not known with certainty. It is thus desirable to relax the recessive and dominant constraint in (2.1), and resort to an optimisation involving the larger penetrance space under the prevalence constraint, which would be numerically more difficult. The poor performance of MALOD under locus heterogeneity is likely due to the fact that penetrances and admixture parameter are confounded in small pedigrees, an extra degree of freedom is unnecessary.

A possible limitation of the power comparisons performed here is that we assumed that all the test statistics conformed to their asymptotic distributions under the null hypothesis. If this is not true and some of the test statistics give anticonservative significance level when asymptotic distributions are used, then the power comparison may unfairly favour these tests. Indeed, we showed that LOD, ALOD and the two NPL statistics are somewhat anticonservative.

In principle this work can be generalised to more complicated family types and disease models but then a complete list of all possible genotype configurations will be computer-intensive since it could be very large. The generalisation of findings with respect to multiplicative models remains to be explored. In this study, a single fully informative marker has been used to focus on different maximisation schemes, which enables exact calculation of the distribution of the test statistics and to some extent approximate multipoint analysis but will lead to two consequences. First, the power will be lower for marker with incomplete information. However, our main concern is the relative performance of these tests, we do not expect it change dramatically. Second, the direct extension of the finding from two-point to multipoint analysis is problematic (Risch & Giuffra 1992) since recombination ceases to be an effective parameter. Moreover, since NPL statistics assess the evidence of allele sharing among affected relative pairs at a test position are more suitable for multipoint analysis.

Other assumptions of this study include HWE of founder trait and marker phenotypes, no interference, ascertainment/selection of small families (Badner et al. 1998, 1999).

2.5 Bibliographic notes

Risch (1989) showed that for simple genetic models and nuclear family data, ignoring heterogeneity and calculating the standard lod score tends to be the more powerful choice unless for small linked fraction α, large pedigrees and small recombination fraction. Davis et al. (1996), Davis & Weeks (1997) examined a variety of statistics for linkage analysis with different genetic models and family structures and showed that NPL had lower power compared to other methods when there was heterogeneity in the data and when families were ascertained through two or more affected children. MOD score is also called MMLS (maximising the maximum lod score, or LOD score maximised over disease models) (Greenberg 1989). In Hodge et al. (1997), the lod score was calculated twice, once assuming a simple dominant model and once assuming a simple recessive model. The maximum over the two models was then reported as result and the type I error rate was corrected for multiple testing. Greenberg et al. (1998) found the power of MMLS method to be robust, with its power ratio to that of the simulated model greater than 0.8 over a range of models. Abreu et al. (1999) compared MMLS and NPL. They simulated 100 data sets with 20 families each, using 26 generating models: (1) 4 intermediate models (penetrance of heterozygote between that of the two homozygotes); (2) 6 two-locus additive models; and (3) 16 two-locus heterogeneity models (admixture $\alpha = 1.0,.7,.5$, and .3; $\alpha = 1.0$ replicates simple Mendelian models). For LOD scores, they assumed dominant and recessive inheritance with 50% penetrance and took the higher of the two maximum LOD scores and subtracted 0.3 to correct for multiple tests (MMLS-C). They compared expected maximum LOD scores and power, using MMLS-C and NPL as well as the true model. Since NPL uses only the affected family members, they also performed an affecteds-only analysis using MMLS-C. The MMLS-C was both uniformly more powerful than NPL for most cases they examined, except when linkage information was low, and close to the results for the true model under locus heterogeneity. They still found better power for the MMLS-C compared with NPL in affecteds-only analysis. The results show that use of two simple modes of inheritance at a fixed penetrance can have more power than NPL when the trait mode of inheritance is complex and when there is heterogeneity in the data set. Nyholt (2000) gave a summary of critical values for lod scores obtained under a variety scenarios including MMLS. Further discussion of allele-sharing statistics was given by Shih & Whittemore (2001).

Chapter 3

Case-control allelic association

3.1 Introduction

Allelic association or haplotype analysis of a set of markers is commonly done on samples of unrelated individuals from one or more populations (Hawley & Kidd 1995; Kidd et al. 2000; Schneider et al. 1997; Terwilliger & Ott 1994; Cox 1998; Zhao, Curtis & Sham 2000). These are also used to study associations between a putative disease locus and marker(s) in samples of cases and controls (Ott 1999; Zhao, Curtis & Sham 2000).

When a single marker is involved, allele frequency differences between cases and controls can be detected using a Pearson contingency table χ^2 (Woolf 1955; Workman & Niswander 1970; Sham & Curtis 1995; Sasieni 1997; Hirotsu et al. 2001) or a log-likelihood ratio statistic. When more markers are involved, it is more appropriate to use a likelihood framework based on haplotype frequencies. The likelihood of observing a given sample is expressed as a function of unknown haplotype frequencies, to be maximised by numerical method such as the EM (Expectation-Maximisation) algorithm. This is called gene counting since the EM steps only involve counting genes. Some simple scenarios including a small number of markers are illustrated in Appendix A. The general situation of a putative disease locus and multiple multiallelic markers has been implemented in early version of **EH** (Xie & Ott 1993; Terwilliger & Ott 1994) and its successor **fehp** (Zhao & Sham 2002).

For complex traits with uncertain disease model, it is natural to adapt model-free statistics similar to those used in linkage analysis. Assuming a generalised single locus model, its parameters can be constrained to yield the same population disease prevalence (K) as considered by Curtis & Sham (1995). The parameters of the disease model are treated as

nuisance parameters in the likelihood function, and a likelihood ratio test is constructed in which the likelihood under the null hypothesis is maximised over only the nuisance parameters, while the likelihood under the alternative hypothesis is maximised over both the nuisance parameters and the association parameters. Similarly the heterogeneity statistic in linkage analysis can be obtained from separate marker-marker analyses of cases only, controls only and combined case-control data.

It is then necessary to address issues concerning about validity and power of such statistics. In general, with its simple, well-known, practical and economic advantages compared to others it is of great interest to fully characterise and apply such design while avoiding spurious association (Long & Longley 1999; Risch & Teng 1998; Devlin & Roeder 1999; Risch 2000). A power analysis using the approach of Long et al. (1997) will be given in Chapter 7.

Chapter aims

This chapter presents a power study of model-free tests in case-control association analysis. Similar study has been reported in Zhao & Sham (1997), Zhao, Curtis & Sham (2000) but current investigation includes ordinary contingency table χ^2 and likelihood ratio tests.

3.2 Methods

Following Xie & Ott (1993), we start by assuming a generalised single locus model, in which a disease locus with one disease allele and one normal allele. Under this model, the allele frequency of the disease allele q and penetrance f_i, $i = 0, 1, 2$ for genotype with i disease allele(s) are similarly defined as in Chapter 2. Contingency table Pearson χ^2 (P) and LRT (G) statistics are also used for comparison. We consider also a haplotype model involving a multiallelic marker B with n ($n \geq 2$) alleles $B_1, B_2, \ldots, B_n$ and allele frequencies $p_1, p_2, \ldots, p_n$. The genotypic probabilities, based on a subject being a case (A) or a control (U), are as follows,

$$\begin{aligned} P(B_iB_i|A) &= \frac{f_0 h_{1i}^2 + f_1(2h_{1i}h_{2i}) + f_2 h_{2i}^2}{K} \\ P(B_iB_j|A) &= \frac{f_0(2h_{1i}h_{1j}) + f_1(2h_{1i}h_{2j} + 2h_{1j}h_{2i}) + f_2(2h_{2i}h_{2j})}{K} \end{aligned}$$

$$
\begin{aligned}
P(B_iB_i|U) &= \frac{s_0 h_{1i}^2 + s_1(2h_{1i}h_{2i}) + s_2 h_{2i}^2}{1-K} \\
P(B_iB_j|U) &= \frac{s_0(2h_{1i}h_{1j}) + s_1(2h_{1i}h_{2j} + 2h_{1j}h_{2i}) + s_2(2h_{2i}h_{2j})}{1-K}
\end{aligned} \tag{3.1}
$$

where $h_{11}, h_{12}, \ldots, h_{1n}, h_{21}, h_{22}, \ldots, h_{2n}$ are the haplotype frequencies, and $s_i = 1 - f_i$. The likelihood of the data is a product of these probabilities, and the parameters of this likelihood function are therefore the disease model parameters and the haplotype frequencies (Sham 1998, p159). The maximum log-likelihood under both linkage equilibrium and linkage disequilibrium are obtained from **fehp** (Zhao & Sham 2002). Since these constitute nested models, twice the difference in log-likelihood should be asymptotically χ^2 with $n-1$ degrees of freedom. Four log-likelihood ratio test (LRT) χ^2 statistics, a heterogeneity statistic, contingency table χ^2 and log-likelihood ratio statistic are considered (Table 3.1).

Table 3.1: Description of test statistics

Statistics	Descriptions	q	f_0	f_1	f_2
Model-based					
T	True model	q, f_0, f_1, f_2 are user specified			
R	Mendelian recessive model	$\sqrt{K}$	0	0	1
D	Mendelian dominant model	$1-\sqrt{1-K}$	0	1	1
F	Maximised over f_1 given K (see description below)	q	$f_0 = f_1$ or $f_1 = f_2$		
Model-free					
H	Heterogeneity statistic	model not needed			
P	Pearson contingency table χ^2	model not needed			
G	LRT χ^2 for contingency table	model not needed			

These statistics are briefly described as follows.

T: This is the statistic one would obtain under a user-specified model, which yields the correct population disease prevalence.

R: This is obtained with disease model parameters being set at Mendelian recessive values ($q = \sqrt{K},\ f_0 = f_1 = 0,\ f_2 = 1$) in the case-control option of **fehp**.

D: This is obtained with disease model parameters being set at Mendelian dominant values Mendelian penetrances ($1 - \sqrt{1-K},\ f_0 = 0,\ f_1 = f_2 = 1$) in the case-control option of **fehp**.

F: This is obtained with disease model parameters being treated as nuisance parameters. The maximum likelihood is obtained from multiple runs of **fehp** with models that are

constrained to produce the correct population risk K similar to **MFLINK** (Curtis & Sham 1995, see also Chapter 2), which implies that only f_1 is free, and that $f_0 = f_1$, $f_2 = 1 - f_1(1-K)/K$ when $f_1 \leq K$, $f_2 = f_1$, $f_0 = (1-f_1)K/(1-K)$ when $f_1 > K$. Specifically, f_1 is restricted to vary along the sides of the triangle of $(0,0,1)$, (K,K,K) and $(0,1,1)$. Given f_0, f_1, f_2, q is obtained by solving $q^2 f_2 + 2q(1-q)f_1 + (1-q)^2 f_0 = K$, namely

$$q = \frac{-2(f_1 - f_0) + \sqrt{[2(f_1 - f_0)]^2 - 4(f_0 - 2f_1 + f_2)(f_0 - K)}}{2(f_0 - 2f_1 + f_2)}$$

and $q = 0.5(K - f_0)/(f_1 - f_0)$ if $(f_0 - 2f_1 + f_2) = 0$ and 1 if $f_2 = K$. Unlike **MFLOD**, this statistic does not maximise f_1 in the denominator of the likelihood ratio, the likelihood under the null is always the likelihood assuming linkage equilibrium. To allow for an extra degree of freedom maximising f_1, this should have n degree of freedom.

H: This is a heterogeneity test in allele frequencies between cases and controls. As described above, the **fehp** program is used three times, once for cases alone, controls alone, and once for the cases and controls pooled together. In each analysis, allele frequencies are estimated and the maximum log-likelihood calculated. The likelihood of an individual is simply the probability of the genotype assuming HWE. Denoting these maximum log-likelihoods as l_{case}, $l_{control}$ and $l_{combine}$, the test statistics -2$(l_{case} + l_{control} - l_{combine})$ is asymptotically χ^2 with $[(n-1) + (n-1) + (n-1)] = n - 1$ degree of freedom. The EM algorithm requires at least two loci, but now only one marker is involved, so a monomorphic marker is used for these individual marker-marker analyses. This statistic tests allele frequency differences between cases and controls assuming Hardy-Weinberg equilibrium.

Statistics **G** and **P** are based on $2 \times n$ contingency table in this case; both are asymptotically χ^2 with $n - 1$ degrees of freedom. Pearson χ^2 statistic is more familiar while likelihood ratio test statistic is more comparable to the model-based statistics above. Note both G and P use alleles rather than genotypes, so they require HWE to be valid tests of association (Sasieni 1997). P and G can also be based on the genotypes but their degrees of freedom will be much larger for comparison with statistics given above. Hence subsequent calculations are based on allele counts.

Haplotype frequencies from Oudet et al. (1993) on fragile X Syndrome are used to evaluate power of model-free tests against correctly or incorrectly specified parametric tests. The frequencies of the seven alleles at DXS548 are 0, 42, 32, 1, 1, 29, 1 on fragile X

chromosomes and 2, 117, 23, 1, 1, 15, 2 on normal chromosomes. The frequency of the disease allele is set arbitrarily to be 0.001 for both R and D. For each single gene disease models given Table 3.2, 500 replicate samples of 1,000 subjects (500 cases and 500 controls) and 500 replicate samples of 10,000 subjects (5,000 cases and 5,000 controls), are simulated in order to investigate the accuracy of the asymptotic χ^2 distribution as a function of sample size.

Five Mendelian models (Table 3.2) with reduced penetrances and phenocopies are used to simulate data in order to examine the performances of test statistics. They vary from simple Mendelian to recessive and dominant models, both rare and common. The choice of these models follows similar argument in Chapter 2. A minor gene model specifies a multiplicative model with genotype relative risk of 4, as has been used earlier for Alzheimer's (Sham & Curtis 1995a, see also Chapter 2).

Table 3.2: The five genetic models

Model	f_0	f_1	f_2	q	K
0 Null (H_0)	0.5	0.5	0.5	0.5000	0.500
1 Rare recessive (RR)	0	0	1	0.0316	0.001
2 Rare dominant (RD)	0	1	1	0.0005	0.001
3 Common recessive (CR)	0.005	0.005	0.5	0.1000	0.010
4 Common dominant (CD)	0.005	0.500	0.5	0.0050	0.010
5 Minor gene (MG)	0.050	0.200	0.8	0.1300	0.100

The properties of each statistic under each model are investigated using simulated samples. The mean value of the test statistics of the replicates is an estimator of its theoretical expectation (which is the sum of the noncentrality parameter and degree of freedom). This allows the noncentrality parameter to be estimated. Under the null hypothesis (H_0), the noncentrality parameter should be 0. A value greater than 0 implies an increase in the false positive rate, while a value less than 0 indicates that the test is conservative. Under an alternative hypothesis, the non-centrality parameter determines the power of the test. At 5% significance level, the values of non-centrality parameter required for 90% power are 17.4 and 18.3 for degrees of freedom 6 and 7, respectively. The required sample size can be extrapolated from the estimate of noncentrality parameter of the simulated samples for any desired level of power. Here the empirical means obtained from simulated samples of size 10,000 are used as approximate estimates of noncentrality parameters so the required

samples sizes can be obtained (90% power at 5% significance, assuming equal number of cases and controls). An exception is specifying null model (H_0) as in Table 3.2 would result in zero log-likelihood.

3.3 Results

The results for 500 replicates of 500 cases and 500 controls are given in Table 3.3. Under H_0, the empirical means exceed the theoretical means (6) for R, D, and H, the reverse is true for F (7). Interestingly, H and G are very close to R. While P and G compare fairly well, the mean of P as a χ^2 statistic is fairly close to its asymptotic value (6) but G tends to be larger. The likelihood-based statistics R, D, H are at least as good as ordinary log-likelihood ratio statistic G, so their inflation under H_0 is acceptable considering the wide use of G. The statistics under the remaining models provide estimates of noncentrality parameters of the associated χ^2 distribtion, which can be used for obtaining sample size given certain significance level and power (data not shown). We expect that under alternative hypotheses the empirical means from simulated samples of size 10,000 are approximately 10 times the corresponding values from that of size 1,000, as demonstrated below.

Table 3.3: Mean and standard deviation of χ^2 statistics from 1,000 subjects (500 replicates of 500 cases and 500 controls)

Model	T	R	D	F	H	P	G
H_0		6.15	6.08	6.38	6.17	6.03	6.18
		3.33	3.22	3.42	3.37	3.22	3.38
RR	327.22	327.22	242.03	327.30	327.22	312.99	327.17
	33.96	33.96	23.16	33.98	33.96	31.83	33.92
RD	109.64	96.63	109.64	109.79	96.63	95.05	96.64
	19.16	18.02	19.16	19.23	18.02	17.46	18.03
CR	85.21	77.09	59.31	86.31	77.09	76.04	77.08
	18.62	17.87	13.34	18.81	17.87	17.40	17.87
CD	30.57	30.37	31.40	32.02	30.37	30.05	30.37
	10.17	10.21	10.32	10.43	10.20	10.04	10.20
MG	31.88	32.88	31.94	33.84	32.90	32.56	32.90
	10.88	11.14	10.73	11.30	11.13	10.98	11.14

The results for 500 replicates of 5,000 cases and 5,000 controls are given in Table 3.4. Under H_0, the empirical means are slightly closer to their theoretical values with the exception of F, which has an empirical mean of 6.24 while the theoretical mean is 7. This

indicates that the asymptotic distribution of this statistic is closer to χ^2 with 6 degrees of freedom than χ^2 with 7 degrees of freedom. Again H and G are fairly close.

Table 3.4: Mean and standard deviation of χ^2 statistics from 10,000 subjects (500 replicates of 5,000 cases and 5,000 controls)

Model	T	R	D	F	H	P	G
H_0		6.15	6.16	6.24	6.16	6.14	6.16
		3.55	3.56	3.60	3.55	3.53	3.55
RR	3199.29	3199.29	2356.60	3199.29	3199.28	3063.71	3198.00
	111.16	111.11	77.51	111.11	111.12	104.43	110.95
RD	1041.95	907.50	1041.95	1041.95	907.50	895.77	907.54
	57.35	52.20	57.35	57.36	52.21	50.89	52.22
CR	804.95	716.72	541.28	805.37	716.69	710.20	716.74
	60.35	57.13	42.14	60.39	57.13	56.10	57.13
CD	261.98	252.31	262.59	262.73	252.31	251.13	252.35
	31.54	30.54	31.63	31.62	30.53	30.25	30.53
MG	268.79	270.35	261.75	270.98	270.23	269.16	270.26
	32.16	32.49	31.32	32.45	32.49	32.23	32.50

The required sample sizes based on 500 replicates of 5,000 cases and 5,000 controls are given in Table 3.5 (by a SAS program in Appendix). It is clear that rare disorders, RR in particular, require less subjects than common disorders. The number of subjects as required by CD is several folds that of CR. The number of subjects as required by major gene model is much larger than that of rare disorder and somewhat between CR and CD considered here.

Table 3.5: Estimated sample sizes required for 90% power and significance level .05 (equal number of cases and controls)

Model	T	R	D	F	H	P	G
RR	55	55	74	58	55	57	55
RD	167	192	167	176	192	195	192
CR	217	243	322	228	243	246	243
CD	665	690	663	697	690	693	690
MG	648	644	665	676	644	647	644

No single test is uniformly more powerful over all 5 models. T has the best performance overall. The power of R and H appear to be equivalent. F is more powerful than H in some situations. For minor gene model, H appears to be substantially more powerful than F. If F is considered to have a χ^2 distribution with 6 degrees of freedom, then F and H would be

equally powerful even in this situation, but then F would be slightly anti-conservative.

This shows that, when the disease model is unknown, the standard χ^2 test of homogeneity of allele frequencies (H) provides nearly optimal power, especially for a minor gene model. If the null distribution of the parametric "model-free" test (F) is accurately known, then this may be preferred to H. Overall, the preferred test for routine use on case-control data is the standard χ^2 test of homogeneity of allele frequencies. For data involving related individuals, however, it is likely that parametric "model-free" methods will have a greater degree of superiority over non-parametric methods.

A final observation is that for R and D models setting allele frequencies to be 0.001 other than obtaining from the correct population prevalence gives similar results (data not shown).

3.4 Discussion

The performance of model-free statistics in case-control allelic association analysis is investigated. It is shown that the heterogeneity statistic is almost as powerful as the likelihood ratio statistic taking true parameter values of a generalised single locus model.

In view of the performance of the six statistics under the null hypothesis, the power comparison is only considered as an approximation. For instance, standard errors of these statistics from Table 4 were about 0.05 ([standard deviation]/$\sqrt{5000}$), the means of R, D, H are greater than 6, while that of F is smaller than 7, the power calcualtion according to noncentral χ^2 would be somewhat inflated for R, D, H but deflated for F, therefore the power comparison may seem difficult. However, similiar behaviour of the popular statistic G offers some reassurance, as their relative performance, except F perhaps. The limitaion of this study is its assumption of single locus disease models, which are unrealistic for complex diseases. It further assumes the absence of hidden population stratification, a potential problem in real application. While the study relies on asymptotic distribution theory, in practice Monte Carlo method should be used whenever appropriate (Zaykin et al. 1995; Lazzeroni & Lange 1997; Zhao, Curtis & Sham 2000).

By including extra markers it is possible to construct a Pearson χ^2 statistic and its likelihood ratio counterpart based on the haplotype table and estimate a global LD statistic, but the Pearson χ^2 statistic may often be inflated and have poor asymptotic property. It

is common to construct such statistics based on the genotypes of these markers. Then test of independence entails "other association" than that by the heterogeneity statistic considered here. These statistics for two markers has been implemented for two markers in **ASSOCIATE** (http://linkage.rockefeller.edu) and a more comprehensive program **2LD** (Zapata et al. 2001). The heterogeneity statistic has been adopted both in the latest version of **EH** and Fallin et al. (2001). There would be concern over the validity of the chi-squared statistics when the number of alleles become large, then empirical significance should be used (Zhao, Curtis & Sham 2000). The concern over use of allelewise versus genotypewise analysis has been extensively discussed by Sasieni (1997).

Both Koch et al. (2000) and Fallin et al. (2001) showed that an association can be detected using either SNPs alone or combined with microsatellite markers surrounding the functional locus even if the functional locus is not typed. It is remarkable that significant results were obtained for haplotypes defined by loci that did not show single-locus significance. Similar remark was made by Longmate (2001). It is likely that when loci influencing disease have alleles whose impact is on the genotype level, tests using that fact may be more powerful. Moreover, the haplotype method does not necessarily help determine the precise position of functional locus. Similar note concerning fine genetic mapping was given by Chiano & Clayton (1998). When parental phase is known it can be directly used, otherwise the E-M algorithm can be applied. While haplotype analysis is invaluable for study of SNPs, where haplotypes will increase the informativeness, there are theoretical and practical issues when many loci are involved (e.g. Zhao, Curtis & Sham 2000).

A final note is that the case-control design should be valuable in homogeneous or isolated populations (Shifman & Darvasi 2001) and it is gaining more attention (Devlin & Roeder 1999; Risch 2000; Ghosh et al. 2000; Zhao et al. 2000; Seltman et al. 2001).

3.5 Bibliographic notes

Chapman & Meng (1966), Guenther (1977) presented a simple formula to calculate noncentrality parameter for alternative hypothesis of type $h_{ij} = p_i q_j + c_{ij}/\sqrt{n}$, $\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} = 0$, i.e., $\lambda = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2/(p_i q_j) + \sum_{i=1}^{m} c_{i.}^2/p_i + \sum_{j=1}^{n} c_{.j}^2/q_j$, where $c_{i.} = \sum_{j=1}^{n} c_{ij}$, $c_{.j} = \sum_{i=1}^{n} c_{ij}$. For two biallelic markers this is comparable to $\lambda = 2nD^2/(p_1 p_2 q_1 q_2)$ (Weir 1996, p113), where p's and q's are the allele frequencies, which is based on standard error of D.

Feder et al. (1996) and Nielson et al. (1998) suggested examining departure from HWE among affected individuals for fine-mapping of disease. For recessive $(\psi, \psi, 1)$ discrete model, the genotype and allele frequencies in the population are expressed as $P(DD|A) = p_D^2/K$, $P(D|A) = p_D(p_D + \psi p_D)/K$, so that the disequilibrium coefficient becomes $D_{DD} = P_{DD} - p_D^2 = [\psi(1-\psi)p_D^2(1-p_D^2)]/K^2$. The quantity to measure the departure from the Hardy-Weinberg equilibrium is $F_D = (H_o - H_e)/(1 - H_e)$ where H_o and H_e are the observed and expected homozygosities, respectively. It can be rewritten as $F_D = [P_{DD|A} + P_{dd|A} - p_{D|A}^2 - p_{d|A}^2]/(1 - p_D^2 - p_d^2) = [\psi(1-\psi)p_D p_d]/K^2$. Association between the disease-susceptibility allele D and a marker allele B can be expressed as $\Delta_{DB} = h_{DB} - p_D q_B$, where h_{DB} is the frequency of haplotype carrying both alleles A and B and q_B is the frequency of marker allele B. For Hardy-Weinberg disequilibrium, $P_{BB|A} = [(1-\psi)(p_D q_B + \Delta_{DB})^2 + \psi q_B^2]/K$ and $q_{B|A} = [\psi q_B + (1-\psi)p_D(p_D q_B + \Delta_{DB})]/K$ so that the Hardy-Weinberg disequilibrium coefficient among affected individuals is $\Delta_{BB|A} = \psi(1-\psi)\Delta_{DB}^2/K^2$. The Hardy-Weinberg departure of Feder et al. (1996) for the marker locus B is $\psi(1-\psi)\Delta_{DB}^2/(K^2 q_B q_b)$ and $F_B = \Delta^2 F_D$ where $\Delta^2 = \Delta_{DB}^2/p_D p_d q_B q_b$. As Δ_{DB} is a function of $(1-\theta)^{2g}$, where g is the number of generations since the founding of the mutation, F_B decays approximately at a rate of $2g\theta$ with recombination distance θ between D and B. With $q_{B|U} = (1-\psi)[q_B - p_D(p_D q_B + \Delta_{DB})]/(1-K)$, the commonly used LD measure (Bengtsson and Thomson 1981, Lehesjoki et al. 1993), $p_{excess} = (q_{B|A} - q_{B|U})/(1 - q_{B|U})$ is

$$p_{excess} = \frac{(1-\psi)p_D\Delta_{DB}}{K(1-K)[q_B + (1-\psi)p_D\Delta_{DB}/(1-K)]}$$

which is only proportional to Δ_{DB} and less sensitive than F_B. For a general disease model with multiallelic susceptibility locus and a marker locus B with m allele, define f_{rs} to be the "apparent" penetrance of genotype $D_r D_s$, so that $K = \sum_r \sum_s f_{rs} p_r p_s$, the marker allele frequencies among the affecteds and unaffecteds are $q_{i|A} = q_i + \delta_i/K$, $q_{i|U} = q_i - \delta_i/(1-K)$ and

$$p_{excess_i} = \frac{\delta_i}{K(1-K)[(1-q_i) + \delta_i/(1-K)]}$$

where $\delta_i = \sum_r \sum_s p_s f_{rs} \Delta_{ri}$ and $\Delta_{ri} = h_{ri} - p_r q_i$. The Hardy-Weinberg disequilibrium coefficients are $\Delta_{ii|A} = h_{ii|A} - q_{i|A}^2 = (K\delta_{ii} - \delta_i^2)/K^2$, $\Delta_{ij|A} = h_{ij|A} - 2q_{i|A}q_{j|A} = 2(K\delta_{ij} - \delta_j\delta_j)/K^2$ and $\delta_{ij} = \sum_r \sum_s f_{rs}\Delta_{ri}\Delta_{sj}$. A contingency table χ^2 can be built up using n cases and n controls as $\chi_{CC}^2 = 4n\sum_i(\tilde{p}_{i|A} - \tilde{p}_{i|U})^2/(\tilde{p}_{i|A} + \tilde{p}_{i|U})$, where $\tilde{p}_{i|A}$ and $\tilde{p}_{i|U}$ are the

sample frequencies of the marker, χ^2_{CC} has χ^2_{m-1} distribution under the null hypothesis of no association and $\chi^2_{m-1}(\delta_{CC})$ distributed under the alternative, where

$$\begin{aligned}\delta_{CC} &= 4n\sum_i \frac{(q_{i|A}-q_{i|U})^2}{q_{i|A}+q_{i|U}} \\ &= 4n\sum_i \frac{\delta_i^2}{K^2(1-K)^2[2q_i+(1-2K)\delta_i/K(1-K)]}\end{aligned}$$

Similarly a Hardy-Weinberg disequilibrium test χ^2_{HW} with $m(m+1)/2-(m-1)-1 = m(m-1)/2$ degree(s) of freedom can be established. The noncentrality parameters for the two tests δ_{CC} and δ_{HW} are given in Nielson et al (1998). They noted that if the penetrances are regarded as genotypic values, much of the theory can be applied to study of quantitative traits (Nielsen and Weir 1999, 2001).

Zaykin et al. (1995) describe simulation procedure to examine power of haplotype analysis using unrelated individuals. Mckeigue (2000) gave the information loss in case-control design relative to family design. Toivonen et al. (2000) discussed data mining.

Chapter 4

Marker characteristics and fine mapping

4.1 Introduction

The importance of marker characteristics in both linkage and association analysis is now well recognised. For example, when a parent passes a gamete to an offspring, a recombination may occur in the parent only if the parent is heterozygous at two joint loci. Ott (1992) examined the sample size necessary to characterise a marker and the effects of wrongly assuming equal allele frequencies to linkage analysis. Terwilliger et al. (1992) focused on the relative importance of marker heterozygosity and intermarker distance and showed that high marker density may be a good substitute for low marker heterozygosity. Kruglyak (1997) investigated how many polymorphic and densely spaced biallelic markers are needed for extraction of most of the mode of inheritance information as compared to microsatellite markers. Kruglyak (1999) used coalescent theory and computer simulation to examine the extent to which LD can extend in the context of genome screen and using SNPs.

Of particular interets is the advantage and disadvantage of SNPs and dense maps over traditional markers in fine mapping. Ott & Rabinowitz (1997), Chapman & Wijsman (1998) both considered the effect of marker polymorphism on mutation detection assuming single ancestral mutation. Ott & Rabinowitz (1997) found that greater marker heterozygosity results in increased power to detect linkage disequilibrium. A similar conclusion was reached by Chapman & Wijsman (1998), that multiallelic markers always have more power to detect

LD than biallelic markers. It is desirable, however, multiple mutations can be incorporated.

Chapter aims

This chapter uses a simple deterministic model to investigate the effect of marker polymorphism on the power of fine mapping. The simple population genetics model incorporates multiple ancestral mutations. Under this model, the effect of marker polymorphism on the power to detect linkage disequilibrium due to multiple mutations is examined for several scenarios. Some results of this investigation are reported in Sham, Zhao & Curtis (2000).

4.2 Methods

We begin with the simplest assumption that current disease chromosomes are due to multiple ancestral mutations at a single marker locus at generation g. We assume the disease locus is recombination rate θ from the marker. We also assume that the marker locus have n alleles, with frequencies $\pi_j, j = 1, \ldots, n$ in the population, so that there are at most m mutations at an allele and n^m possible patterns of mutations in total. Let $d_i = 1, \ldots, m$ be the specific pattern of each of these n^m enumerations, with a non-zero value indicating a mutation at that allele and their counts accumulating to c_j, so that marker allele frequency at the disease-carrying chromosome at generation 0 is $f_j = c_j/m$, $j = 1, \ldots, n$. Under random mating and nonoverlapping generations, the expected frequencies of the marker alleles among chromosomes containing a disease mutation at generation g are obtained via the following basic relation

$$f_j^{(g)} = f_j(1-\theta)^g + [1-(1-\theta)^g]\,\pi_j \tag{4.1}$$

i.e., the expected frequencies for the marker chromosome at generation g is due to chromosomes that remain intact with the founder mutations and those experienced at least one recombinations. The allele frequencies discrepancy between the disease chromosome and chromosomes in the population can be measured by Pearson χ^2 and likelihood ratio test statistics as follows.

$$\chi_i^2 \;=\; N\sum_j \frac{(\pi_j - f_j^{(g)})^2}{\pi_j + f_j^{(g)}} \tag{4.2}$$

and

$$l_i = 2N \sum_j \left[\pi_j \ln \pi_j + f_j^{(g)} \ln f_j^{(g)} - (\pi_j + f_j^{(g)}) \ln \frac{\pi_j + f_j^{(g)}}{2} \right] \tag{4.3}$$

for $i = 1, \ldots, n^m$ and N is the appropriate sample size. Asymptotically χ_i^2 and l_i would have noncentral χ^2 distribution with degree of freedom $n-1$, and noncentrality parameters χ_i^2 and l_i respectively.

The appropriate area s_i is calculated as the power estimate from these two distributions that exceeds the critical value set by the global α value, the type I error under the null distribution of χ^2. Denote $w_i = \prod_j \pi_j^{c_j}, i = 1, \ldots, n^m; j = 1, \ldots, n$ the multinomial probability associated with the $i-th$ of the n^m enumerations, we can get an overall estimate of the power by $\sum_i w_i s_i$, $i = 1, \ldots, n^m$. The estimate of N can be iteratively refined to achieve predefined power. In practice to enumerate all the n^m possibilities is quite slow so we can set a number of values that are interesting to us and get their overall estimates of power. The minimum and maximum of χ_i^2 and l_i can also be recorded for each mutation. The n^m possibilities are enumerated using a recursive algorithm.

Two sets of allele frequencies are examined for markers with alleles varying from 2 to 15, which represents fairly common polymorphisms. In the first scheme each allele has frequency $\pi_j = 1/n$ for $j = 1, \ldots, n$, so the marker heterozygosity is easily calculated as $1 - \sum \pi_j^2 = 1 - 1/n$. In the second scheme different alleles follow a geometric relationship, $p_1 = (1-r)/(1-r^n), \pi_j = \pi_{j-1} r, j = 2, \ldots, n$, the marker heterozygosity is $1-(1+r^n)/(1+r)p_1 = [2r/(1+r)][(1-r^{(n-1)})/(1-r^n)]$. By setting $r = 0.5$ here allele 1 has the highest frequency 0.67 for biallelic marker and 0.5 for marker with 15 alleles, while the marker heterozygosity increases from 0.44 to 0.67. A proportion of $r = 9/11 \approx 0.82$ is used so that allele 1 has the highest frequency 0.55 for biallelic marker and 0.19 for 15-allele marker while the marker heterozygosity increases from 0.50 to 0.89.

The following combinations are examined for 1 to 5 mutations: the generation $g = 20$, 50 and 80, the recombination fractions $\theta = 0.005$ and 0.05. It is expected that combinations such as $g = 50$, $\theta = 0.005$ and $g = 50$, $\theta = 0.05$ will give comparable results to $g = 500, \theta = 0.0005$ and $g = 500, \theta = 0.005$.

The entire calculation is done by a C program calling up a number of subroutines: subroutine *digit* for the recursive enumeration, Fortran routines to calculate central and

noncentral χ^2 distribution functions and noncentrality parameters, unidimensional optimisation procedures *fibonacci* and *fmin* to determine the required sample size for a given type I error rate and power. Subroutines *digit*, *fibonacci*, *fmin* are listed in Appendix B. Fortran routines are based on algorithms published in Applied Statistics (AS91, AS170, AS245, AS275, available from http://lib.stat.cmu.edu).

4.3 Results

Table 4.1 shows required the sample size based on the exact calculation. In general, the likelihood ratio test statistic exhibits more power (data not shown) than Pearson χ^2. As the number of mutations increases, the effect of number of alleles is more pronounced.

Table 4.1: Required sample sizes for $\alpha = 0.001$ and $1 - \beta = 0.9$ over $g = 20, 50, 80$ and $\theta = 0.005, 0.05$

	20				50				80			
	0.005		0.05		0.005		0.05		0.005		0.05	
mutation=1												
2	41	61	315	521	59	90	7000	12404	83	130	153301	273655
3	30	49	207	396	43	71	4166	9138	59	102	88030	200080
4	27	47	170	372	38	68	3132	8435	51	96	64423	184002
5	26	48	152	372	36	69	2594	8373	48	97	51989	182552
6	26	49	142	380	35	71	2262	8557	46	100	44229	185929
7	25	51	135	393	34	73	2037	8813	45	103	38891	191671
8	25	53	131	407	34	76	1874	9124	45	107	34973	198183
9	26	54	128	421	34	78	1752	9427	45	111	31966	204930
10	26	56	127	435	34	81	1656	9734	45	114	29579	211906
11	26	58	125	448	35	83	1579	10033	45	118	27635	218092
12	26	60	125	461	35	86	1516	10321	46	121	26017	224348
13	27	61	124	473	35	88	1464	10598	46	125	24659	230376
14	27	63	124	485	36	90	1421	10865	46	128	23476	236202
15	27	64	124	497	36	93	1386	11123	47	131	24560	241776

mutation=2

2	a	348	a	500	a	463	a	44655	a	625	a	962221
3	70	105	597	806	103	151	14000	18118	149	213	313316	394666
4	47	84	364	649	68	121	8157	14590	96	171	177318	317607
5	39	80	284	614	56	115	6018	13761	78	162	128958	299360
6	36	80	243	615	50	115	4901	13744	70	162	103561	298950
7	34	82	219	629	47	118	4208	14042	65	166	87774	305250
8	33	85	203	647	45	121	3735	14469	62	171	76781	314506
9	32	87	191	669	44	125	3387	14954	59	177	68822	325265
10	32	90	183	691	43	130	3122	15440	58	183	62571	335538
11	31	93	177	713	43	134	2914	15926	57	188	57733	346085
12	31	96	172	733	42	137	2745	16397	56	194	53700	356640
13	31	98	168	754	42	141	2606	16857	56	199	50318	366325
14	31	101	165	774	42	145	2488	17303	56	204	47487	375945
15	31	103	163	793	42	149	2388	17724	56	209	45054	385233

mutation=3

2	416	b	2704	b	566	b	58884	b	768	b	1278369	b
3	c	159	c	1287	c	229	c	31032	c	325	c	688633
4	90	114	792	908	133	164	19342	21294	193	233	426784	469190
5	60	104	490	821	88	150	11371	19050	125	212	250193	418463
6	50	103	382	806	72	148	8917	18605	101	209	185013	408269
7	45	104	325	817	64	150	6968	18815	89	212	149953	412510
8	42	107	289	838	59	154	5991	19298	81	218	127695	422995
9	40	111	265	864	56	159	5312	19893	76	225	112179	436012
10	38	114	248	892	53	164	4811	20541	73	232	100678	449925
11	37	118	235	919	52	169	4424	21166	70	239	91740	463938
12	37	121	225	946	51	174	4116	21796	69	246	84601	478114
13	36	124	217	973	50	179	3865	22406	67	252	78743	491271
14	36	127	210	998	49	183	3658	23011	66	259	73836	504429

15	36	130	205	1022	49	188	3481	23578	65	265	69670	516839

mutation=4

2	d	1147	d	7608	d	1564	d	168285	d	2131	d	3666858
3	402	349	2549	2066	538	461	51458	43946	721	612	1111038	949769
4	145	149	1126	1140	209	213	25214	25979	296	300	600000	567967
5	94	128	812	987	138	184	19767	22497	199	259	438705	491914
6	69	124	561	953	100	178	13159	21686	143	250	300000	474056
7	58	125	453	958	84	179	10500	21796	118	252	200000	476267
8	52	128	390	980	75	183	8948	22282	105	257	185964	486771
9	48	131	349	1008	69	188	7465	22913	96	265	160620	500750
10	46	135	320	1038	65	194	6668	23618	90	273	142447	516039
11	44	139	299	1069	62	199	6066	24324	85	281	128703	531745
12	43	143	283	1100	60	205	5594	25035	82	289	117840	547029
13	42	147	269	1130	58	210	5213	25723	80	297	109060	562118
14	41	151	259	1158	57	216	4900	26377	78	304	101786	576605
15	40	154	250	1186	56	221	4634	27019	76	311	95649	590692

mutation=5

2	1113	1841	7137	11455	1507	2476	155053	245369	2038	3328	3365250	5310142
3	482	532	3303	3408	663	720	73968	73848	910	975	1616146	1601564
4	370	192	2178	1428	492	305	44962	32560	655	380	964316	713259
5	120	156	1929	1176	173	272	20801	26795	244	311	452149	586413
6	91	147	726	1115	132	222	16362	25343	188	295	355886	554258
7	74	147	500	1112	107	210	14000	25253	153	294	300000	552080
8	64	150	499	1133	93	214	11191	25723	131	300	254320	562118
9	58	154	440	1163	84	219	10500	26395	118	308	200000	577296
10	54	158	398	1197	77	225	8986	27181	108	317	184955	593975
11	51	163	367	1233	73	232	7765	27980	102	326	166406	611709
12	49	167	344	1269	70	239	7121	28783	97	335	151890	629200
13	48	172	325	1303	67	245	6605	29570	93	344	140077	646404

14	46	176	310	1336	65	251	6180	30339	90	353	130371	663337
15	45	180	298	1368	63	257	5824	31072	87	362	122251	679534

The maximum attainable powers for these cases are (a) 0.500, (b) 0.778, (c) 0.556 and (d) 0.625. See text for more details.

Several observations can be made. Firstly, for the three generations considered here the decay of linkage disequilibrium is not so remarkable for $\theta = 0.005$ compared to $\theta = 0.050$. Secondly, power usually increases with the number of alleles, so the sample size required for biallelic marker may be ten times or more for marker with fifteen alleles. The power seems to attain its maximum at about 10 alleles. Finally, a larger sample size is required for more mutations. For $g = 20$ generations this increase is very slow but for $g = 50$ and $\theta = 0.05$, the increase is roughly proportional to number of mutations. Moreover, the log-likelihood ratio statistic provide slightly larger power than Pearson χ^2.

There are four special cases; the mutation-allele combinations 2-2, 3-3 and 4-2 under equal allele frequencies when there is one mutation in each allele, and 3-2 under unequal allele frequencies when there are two mutations occurring in the more frequent allele and one mutation occurring in the less frequent allele. In such cases, the noncentrality parameters reduce to zero, so it is impossible to achieve more power than certain magnitude by the statistics proposed here. Suppose the number of instances with zero noncentrality parameter is n_0, the maximum attainable power is predicted as follows,

$$1 - n_0(1 - \alpha)w_i \tag{4.4}$$

where α is the type I error rate, w_i is the probability weight.

For equifrequent marker, according to equation (4.1), the three combinations each result in the same expected allele frequencies as observed in n_0=2, 6 and 6 instances, respectively. Since the weight w_i is $1/n^m$, the optimal powers are therefore 0.500, 0.778 and 0.625. Under unequal allele frequencies the weight is 4/27 and there are $n_0 = 3$ such instances so the optimal power is 0.556.

4.4 Discussion

Fine mapping of disease mutation depends on both mutation-marker distance and age of mutation. We have in addition quantified the effect of number of mutations and number of alleles. When $\theta \ll 0.5$, equation (4.1) is roughly a linear function of $g\theta$, so that for old mutations to be detected, the marker has to be fairly close. Even in population isolates, recent mutations are less interesting than those brought in by founders. This has important implications with respect to the recent attention to SNPs (Guo & Lange 2000).

We only examined two simple scenarios, one assuming equal allele frequencies and one with a geometric decrease between alleles. Results for 20 and 80 generations not reported in Sham, Zhao & Curtis (2000) are also included. Equal allele frequency tend to have greater power, as noted by Ott (1992) in the context of linkage analysis. The optimal number of alleles won't necessarily be large for detecting specific number of mutations although the reality may be more complicated. We do not consider actual mutation mechanism such as single and multiple step mutations (Farrall & Weeks 1997). Many factors such as selection or population growth rate would affect the result. An obvious note from Chapman & Wijsman (1998) is that when choosing between markers with the same heterozygosity and different number of alleles, the one with fewer alleles will be more powerful.

The current model is simple to interpret and easy to apply to multiple mutations. Ott & Rabinowitz (1997), Chapman & Wijsman (1998) used simple deterministic models and assumed that a single founder of the population was responsible for a mutation at the disease locus at g generations ago, and a marker locus with recombination rate θ with the disease mutation was used, which had n alleles each with frequencies $\pi_1, \ldots, \pi_n$. Ott & Rabinowitz (1997) assumed that a priori any member of the population was equally likely to be the founder so the probability that mutation coupled with i-th allele was π_i. After g generations a proportion $\rho = 1 - (1-\theta)^{(g-1)}$ had undergone at least one recombination between the disease locus and the marker locus. Also among the proportion of chromosomes in which the disease mutation undergone a recombination, the distribution of marker alleles was not different from that of general population, which implied there were random mixing and negligible genetic drift. Let O_i denote the number of chromosomes in the sample with the i-th allele, and $E_i = N\pi_i$ the expectation under the null hypothesis of no disequilibrium, a Pearson χ^2 statistic $\sum_{i=1}^{n} (O_i - E_i)^2/E_i$ can be constructed and referred to χ^2 distribution

with $n-1$ degrees of freedom. If the original mutation was in coupling with a fixed but arbitrary allele $i^\star$, so that the prevalence of the $i^\star$-th allele in the population of chromosomes containing the original mutation was $(1-\rho)+\rho\pi_{i^\star}$, the noncentrality parameter became $N(1-\rho)^2(\pi_{i^\star}^{-1}-1)$, or $N(1-\rho)^2(n-1) \equiv N(n-1)(1-\theta)^{2(g-1)}$ with equal allele frequencies, where N is the sample size. Setting $N=100$ and assume equal allele frequencies, they conducted 3,600 simulations for $g=20,50,80$ and $n=2,\ldots,10$. In Chapman & Wijsman (1998), a mutation introduced into the population had a very small frequency r, then the marker allele frequencies in the disease mutation-carrying chromosomes and the normal chromosomes were

$$x_j^{(g)} = \begin{cases} [1-(1-\theta)^g]\pi_j & j=1,\ldots,i-1,i+1,\ldots,n \\ (1-\theta)^g + [1-(1-\theta)^g]\pi_j & j=i \end{cases}$$

and $y_j^{(g)} \approx \pi_j,\ j=1,\ldots,n$, where i was the index for initial mutation on a background of allele i. They examined several disease models. In the recessive case, the observed marker allele frequencies in cases and controls were roughly $x_j^{(g)}$ and $y_j^{(g)},\ j=1,\ldots,n$; in the dominant case, $0.5(x_j^{(g)}+\pi_j)$ and $y_j^{(g)},\ j=1,\ldots,n$; for mixture model with any other fractions. The same χ^2 is constructed but Chapman & Wijsman (1998) assumed marker polymorphism proceed the disease mutation and an expected χ^2 statistic was obtained from the weighted average of the χ^2 for each allele. Their considerations were mainly for biallelic and multiallelic markers with equifrequent alleles. Equation (4.1) is comparable to Ott & Rabinowitz (1997) for one mutant allele. Since $(1-\theta)^g \approx 1-\theta g$, we have $\rho \approx \theta g$, their basic relation becomes $1-\theta g+\theta g\pi_{i^\star}$. By equation (4.1), it would be $(1-\theta g+\theta g\pi_j)$, and we evaluate w_j by enumeration instead of by simulation. Chapman & Wijsman (1998) also had similar forms.

4.5 Bibliographic notes

Consider a specific haplotype (A_i, B_j) formed by loci A and B whose alleles are indexed by i and j. Suppose that start from generation 0, all individuals are produced by random mating. Then the probability of haplotype (A_i, B_j) in generation 1 is $h_1(A_i,B_j) = (1-\theta)h_0(A_i,B_j) + \theta P(A_i)P(B_j)$, where θ is the recombination rate between A and B, and $h_0(A_i,B_j)$ is the probability of (A_i,B_j) in generation 0. The disequilibrium in generation 1

is $\Delta_1 = h_1(A_i, B_j) - P(A_i)P(B_j) = (1-\theta)\Delta_0$, where Δ_0 is the disequilibrium in generation 0. After g generations, the disequilibrium is $\Delta_g = (1-\theta)^g\Delta_0$. For loci that are unlinked $(1-\theta = 1/2)$, the equilibrium is reached quickly; for example, $(1/2)^{10} = 1/1024$. In contrast, for tightly linked loci the disequilibrium decays very slowly; for example $(1-\theta)^{10} > 1/3$ for $\theta = 0.1$. In the case of tightly linked loci $(1-\theta)^g \approx 1 - g\theta$, and θg dominates the decay of disequilibrium. Similar result can be obtained for multiple loci (Geiringer 1944; Bennett 1954). For simplicity subscripts of loci A, B and C are dropped and the expressions become $h_{g+1}(AB) = (1-\theta)h_g(AB) + \theta P(A)P(B)$, $\Delta_g(AB) = h_g(AB) - P(A)P(B)$, $\Delta_g(AB) = \theta\Delta_{g-1}(AB) = (1-\theta)^g\Delta_0(AB)$. For three loci A, B and C, we have $h_{g+1}(ABC) = (1-\theta_{AB})(1-\theta_{BC})(1-\theta_{AC})h_g(ABC) + \theta_{AB}(1-\theta_{BC})P(A)h_g(BC) + \theta_{AB}(1-\theta_{AC})P(B)h_g(AC) + \theta_{AC}(1-\theta_{AB})P(C)h_g(AB)$, and if we define $\Delta_g(ABC) = h_g(ABC) - P(A)\Delta_g(BC) - P(B)h_g(AC) - P(C)\Delta_g(AB) - P(A)P(B)P(C)$, then $\Delta_g(ABC) = (1-\theta_{AB})(1-\theta_{BC})(1-\theta_{AC})\Delta_{g-1}(ABC) = [(1-\theta_{AB})(1-\theta_{BC})(1-\theta_{AC})]^g\Delta_0(ABC)$.

Hästbacka et al. (1992) noted that the best chance of success with allelic association mapping is when most of the disease chromosomes in the population descend from a single ancestral mutation and the mutation is old enough to allow recombination to break up the ancestral haplotype, but not so old that the maintained neighbourhood around the disease locus is too small to be easily detected. These conditions often hold for isolated founder populations that are not very old. If one can identify an isolated human population in which a significant fraction of affected individuals have inherited the predisposing allele from a common ancestor, one can exploit the tremendous power of LD mapping (allelic association mapping). The resolution of recombinational mapping in a limited family series can be greatly improved, even in the case of relatively recently expanded genes in isolated founder populations, for disease chromosomes that descend from the same ancestral mutation should share a common haplotype in a neighborhood of the disease locus, reflecting the haplotype on the ancestral chromosome on which the mutation occurred. De la Chapelle (1993, 1998) gave many practical examples.

Several methods were proposed to formulate expected haplotype frequencies assuming affected individuals are descended from a common founder, especially for Mendelian disorders in population isolates such as Finland (Terwilliger 1995; Jorde 1995; Kaplan 1995; Devlin et al. 1996; Xiong & Guo 1997; Guo 1997; Rannala & Slatkin 1998; Lazzeroni 1998). These developments rely on simplicity of the underlying population genetics model. The

methods of both Kaplan et al. (1995) and Devlin et al. (1996) would only be effective when a single haplotype currently has a high frequency in disease chromosomes. Terwilliger's method combines the likelihood ratio statistics from separate LD analyses of each marker and does not use all the information in a haplotype. Service et al. (1999) evaluates a LD-mapping method for genome screening in founder populations which make direct use of haplotype information assuming affected individuals are related to a common founder. Rannala & Slatkin (1998) showed this method will produce maximum likelihood estimate when genealogy of affected individuals is of "star" type, which implies individuals are related only through one common ancestor. Other works include Thompson & Neel (1997), Lazzeroni (1998), Graham & Thompson (1998, 2000), McPeek & Strahs (1999), MacLean et al. (2000), Morris et al. (2000), Wright et al. (2000). Clayton (2000) & Lazzeroni (2001) gave comprehensive reviews of these works. A coalescent approach to study LD between SNPs was also conducted by Zöllner & von Haeseler (2000). A Bayesian approach has been adopted by Liu et al (2001).

A procedure developed by Waller et al. (1995) for evaluating power is described here. Consider the distribution function of $Z = \sum_{k=1}^{n} a_k Y_k$, where $a_k, k = 1, \ldots, n$ are the weights taking real values, and $Y_k, k = 1, \ldots, n$ are independent non-central chi-squared random variables with n_k degrees of freedom and noncentral parameters δ_k, for $k = 1, \ldots, n$. The distribution functions of Y_k and Z do not have closed form, but their characteristic functions do. From above the characteristic function for Z is $\phi_Z(t) = \prod_{k=1}^{n} \phi_{Y_k}(a_k t)$. Also $E(Z) = \mu = \sum_{k=1}^{n} a_k(n_k + \delta_k)$ and $V(Z) = \sigma^2 = \sum_{k=1}^{n} a_k^2(n_k + 2\delta_k)$. We scale Z to obtain $Z' = (Z - \mu)/\sigma$, whose characteristic function is given by the formula $\phi_{Z'}(t) = \exp(-it\mu)\phi_Z(t/\sigma)$, the following inversion formula can be utilized to obtain the cumulative distribution function.

$$F_Z(z) = \frac{1}{2} + \frac{\mu z}{2\pi} - \sum_{n=1-H}^{H-1} \frac{\phi_Z(\eta n)}{2\pi i n} e^{-i\eta n z}, \quad n \neq 0$$

where η is chosen to ensure the range of $F_Z(z)$ is fully represented, i.e., the values of $F_Z(z)$ include both 0 and 1. H is chosen to define the $(2H-1)$ points where the cumulative distribution function is estimated. The values for z may be chosen as the Fourier frequencies, that is $z_j = (2\pi(j - H)/2\eta(H - 1)))$ for $j = 1, \ldots, 2H - 1$. Thus the sum can be efficiently calculated using FFT.

Chapter 5

Computer simulation methods

5.1 Introduction

Common diseases such as coronary heart disease have a complicated aetiology which involves, among other things, genetic heterogeneity, epistasis, pleiotropy and gene-environmental interaction, which underlies most theoretical explorations. For example, the RFA proposal for GAW 12 problem II delineated such a system in which gene and environmental effects operate together on disease susceptibility, and the major etiological components underlying the system are simplified into five individual quantitative traits. Nevertheless a particular feature often overlooked by most such computer programs is LD.

Computer simulation can also be used for practical problems. Perhaps the most established application is power analysis of a linkage study via computer programs **SLINK** and **SIMLINK**. Given certain family structures and genetic markers, simulation-based statistics are obtained by generating replicates of the observed families structures conditioning on available phenotype information. Assuming that each replicate occurs with equal probability, expected lod score is simply the average of lod scores of all replicates. This quantity at the true recombination value is designated as ELOD. The maximum of expected lod score (MELOD) over different recombination values and the expected maximum lod score (EMLOD) can also be obtained. MELOD may not be ELOD if the estimated recombination value is inconsistent due to sampling schemes or the wrong assumption of parameter values. EMLOD is an average of the maximum lod scores obtained in each replicate. EMLOD of a pedigree conditional on the heterozygosity/homozygosity status of each pedigree member provides means of identifying pedigree member(s) whose marker status have a strong impact

on the linkage information provided by the pedigree, which yields a large difference in the EMLOD for his/her pedigree depending on whether or not (s)he is marker heterozygous. By definition ELOD $\leq$ MELOD $\leq$ EMLOD (Terwilliger & Ott 1994; Ott 1999).

Chapter aims

This Chapter highlights two scenarios of computer simulation. First, a general procedure for simulating oligogenic trait, taking into account LD, is introduced. The disease model incorporates major gene effects as well as polygenic and environmental effects with genetic loci in linkage equilibrium or disequilibrium. This has been used to study power for a linkage disequilibrium problem (unpublished data). Second, statistical power analysis using computer simulation is applied to two examples of homozygosity mapping (Smith 1953; Lander & Botstein 1987).

5.2 Methods, implementations and results

Simulation of oligogenic traits

A genetic model incorporating major gene, polygenic and environmental effects is as follows:

$$\mathbf{x} = \mathbf{g} + \mathbf{p} + \mathbf{c} + \mathbf{e} \tag{5.1}$$

where $\mathbf{x}$ is a multivariate trait $(x_i, i = 1, ..., NTRAIT)$, a result of additive effects of major loci $\mathbf{g}$ $(g_j, j = 1, ..., NMG)$, polygenic effects $\mathbf{p}$ $(p_k, k = 1, ..., NPG)$, common environment $\mathbf{c}$ $(c_l, l = 1, ..., NCE)$ and unique environment $\mathbf{e}$ $(e_m, m = 1, ..., NUE)$. The variables $\mathbf{g}, \mathbf{p}, \mathbf{c}, \mathbf{e}$ are collectively referred to as causal components of the trait. The NMG major loci are a subset of $NLOCI$ loci for which the order of loci, their recombination fractions $(\theta_1, \theta_2, ..., \theta_{NLOCI-1})$, the number of alleles and allele frequencies of them are specified. The meaning of some model constants are summarised in Table 5.1.

The polygenic effect of an offspring, conditional on those of parents (p_F, p_M), is $p_o = (p_F + p_M)/2 + u/\sqrt{2}$, $u \sim N(0, 1)$ is specific for each offspring. Suppose the degree of transmission of parental shared environment to offspring is k, then under no assortative mating, the common environment for offspring is $k(c_F + c_M) + v\sqrt{(1 - 2k^2)}$, where $v \sim N(0, 1)$ is the same within the whole sibship. The valid range for k is $(0, 1/\sqrt{2})$. The

Table 5.1: Some model constants

Constants	meaning
$MAXFAM$	maximum # of families
$MAXIND$	maximum # of individuals within a family
$MAXLOCI$	maximum # of loci
$MAXALLELES$	maximum # of alleles at each locus
$NTRAIT$	# of traits
$NLOCI$	# of loci
NMG	# of major genes
NPG	# of polygenes
NCE	# of common environments
NUE	# of unique environments
β	matrix of regression coefficients for each trait

unique environment is different among individuals and is $N(0,1)$. All latent variables ($\mathbf{g}$, $\mathbf{p}$, $\mathbf{c}$, $\mathbf{e}$) have mean 0 and variance 1.

The regression equation of trait x on latent variables is

$$\begin{aligned} x_i &= \beta_{ig_1} g_1 + \cdots + \beta_{ig_{NMG}} g_{NMG} + \beta_{ip_1} p_1 + \cdots + \beta_{ip_{NPG}} p_{NPG} \\ &+ \beta_{ic_1} c_1 + \cdots + \beta_{ic_{NCE}} c_{NCE} + \beta_{ie_1} e_1 + \cdots + \beta_{ie_{NUE}} e_{NUE} \end{aligned} \tag{5.2}$$

The mean and variance of x_i are then 0 and $\sum \beta_{ij}^2, j = 1, ..., NMG + NPG + NCE + NUE$. To ease the specifications, the input β's can be standardised by this to yield unit trait variance.

It is possible to generate disequilibrium between multiallelic markers systematically. Without loss of generality, consider two marker loci, with alleles m and n, allele frequencies $p_1, p_2, ..., p_m,\ q_1, q_2, ..., q_n$, we can then arrange the $m \times n$ haplotype frequencies into a contingency table, with some cells taking values specfied by the user. The $m \times n$ contingency table χ^2 statistic becomes $\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} h_{ij}^2 / p_i q_j - 1$. The χ^2 statistic can be rewritten as a familiar quadratic form. let $\mathbf{w}$ be the $m \times n$ diagonal matrix of direct product of the table marginals, the χ^2-squared statistic is $\mathbf{h}\mathbf{w}^{-1}\mathbf{h}' - 1$, where $\mathbf{h}$ contains the actual haplotype frequencies. Under Hardy-Weinberg equilibrium, the haplotype frequencies is obtained via numerical optimisation in a SAS program, with the constraints of haplotype frequency estimates being nonnegative. For $m = n = 2$, the

haplotype frequencies can be parameterised by allele frequencies and a single linkage disequilibrium parameter D, to be $p_1q_1 + D$, $p_1q_2 - D$, $p_2q_1 - D$ and $p_2q_2 + D$. The constraints are $\max(-p_1q_1, -p_2q_2) = -\min(p_1q_1, p_2q_2) \leq D \leq \min(p_1q_2, p_2q_1)$, $p_1q_1 + D \leq p_1$ and $p_2q_2 + D \leq q_2$, $D \leq \min(p_2q_1, p_1q_2)$ to ensure nonnegative cell and marginal probabilities. In summary, -$\min(p_1q_1, p_2q_2) \leq D \leq \min(p_1q_2, p_2q_1)$. With another biallelic marker in equilibrium with the second marker, with allele frequencies r_1, r_2, the haplotype frequencies will be $(p_1q_1 + d)r_1$, $(p_1q_1 + d)r_2$, $(p_1q_2 - d)r_1$, $(p_1q_2 - d)r_2$, $(p_2q_1 - d)r_1$, $(p_2q_1 - d)r_2$, $(p_2q_2 + d)r_1$ and $(p_2q_2 + d)r_2$. We first get parental haplotypes from population frequencies and then proceed to generate the genotypes for children. Some loci are assigned as major loci with dominance parameters. The disequilibrium is simulated if the disequilibrium matrix of haplotype frequencies is specified.

The implementation is made compatible with pedigree files similar to **LINKAGE** and **SIMULATE**. Details are givein in Figure 5-1.

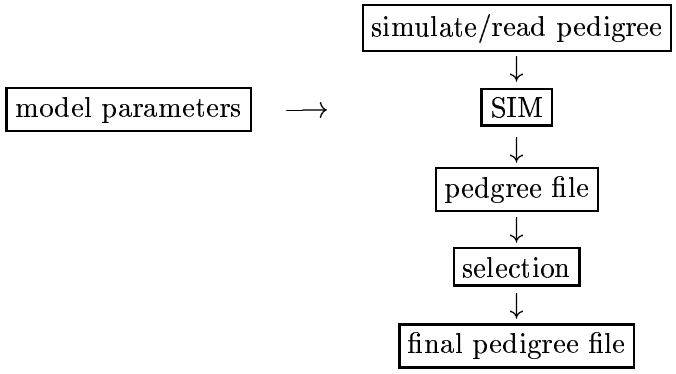

Figure 5-1: Flowchart of SIM

Step SIM processes one family at a time, so that certain selection procedure can be adopted easily. For simulation of genotypes, the usual way of obtaining allele in a specific locus was described in Weir (1990). Consider a 4-allele locus with allele frequencies p_1, p_2, p_3, p_4, when drawing a (pseudo)random number from [0,1] we could divide the unit interval into four segments, with boundaries $0, p_1, p_1 + p_2, p_1 + p_2 + p_3, 1$, these segments therefore have lengths p_1, p_2, p_3, p_4. We assign allele number i if the random number falls within the ith interval. Similar logic is applicable to simulation of recombination events. A comprehensive C library, RANLIB, is used for generating random variables. Numerical Recipes (Press et al. 1992) utility nrutil.h/nrutil.c is used to create dynamic arrays

and matrices. A compilation control file Makefile has been created and tested under Sun Solaris and DEC Alpha with GNU C compiler, PC with Symantec C++ and Cygwin. Prolix output could be obtained by commenting on #undef DEBUG statement in include/sim.h. A program **diseq** is provided to simulate nuclear families assuming linkage disequilibrium. The sibship size distribution is based on the popular negative binomial distribution or truncated Poisson distribution, i.e., a particular class cannot be observed and eliminated from the sample space (Casella & Berger 1990, p123; Sham et al. 1997). Thus $P(X_T = x) = P(X = x)/P(X > 0),\ x = 1, 2, \ldots,$ is the truncated distribution where class 0 cannot be observed. With Poisson distribution we have $P(X = 0) = \mu$, the truncated distribution is $P(X_T = x) = \mu^x e^{-\mu}/x!(1 - \mu),\ x = 1, 2, \ldots.$

A version of **CHRSIM**, a chromosome-based simulation program, **SIMULATE**, a program for data simulation under no linkage, together with the **LINKAGE** utility routine makeped.c have been included as separate programs. The method in **CHRSIM** has been extended to allow inclusion of random markers based on recent estimates of chromosome lengths and Genethon map (Dib et al. 1996, ftp://ftp.genethon.fr/pub/Gmap/Nature-1995/), which would make it more flexible and suitable for genome scan research.

Homozygosity mapping

Familial hemophagocylic lymphohistiocytosis (FHL), also known as familial erythrophagocytic lymphohistiocytosis and familial histiocytic reticulosis, is a rare Mendelian recessive disorder of early childhood characterised by excessive immune activation. The population prevalence of the disorder is approximately 1/50000, so that the disease allele frequency would be 0.00472. It is that expected markers with 5 equal frequent alleles and heterozygousity of 80% is a good approximation. There are five families available (see Figures 5-2-Figure 5-6). The availability of genotyping is 4, 7-9, 12-17 for family 1, 7-8, 11-20 for family 2 8-13 for family 3 17-26, 29-32, 36-39 for family 4 and 18-19 for family 5.

Surprisingly with such an ordinary simulation problem, it took 58 hours on DEC Alpha to get a replicate using **SLINK**. So only 100 replicates would take almost a year for data simulation! To get around this the **SIMULATE** is used to get many replicates, family by family, keeping those replicates whose phenotypes of each family member matched the observed phenotypes (**SIMULATE** simulates under no linkage without conditioning on the phenotypic information of other family members, so genotypes of three loci are simulated

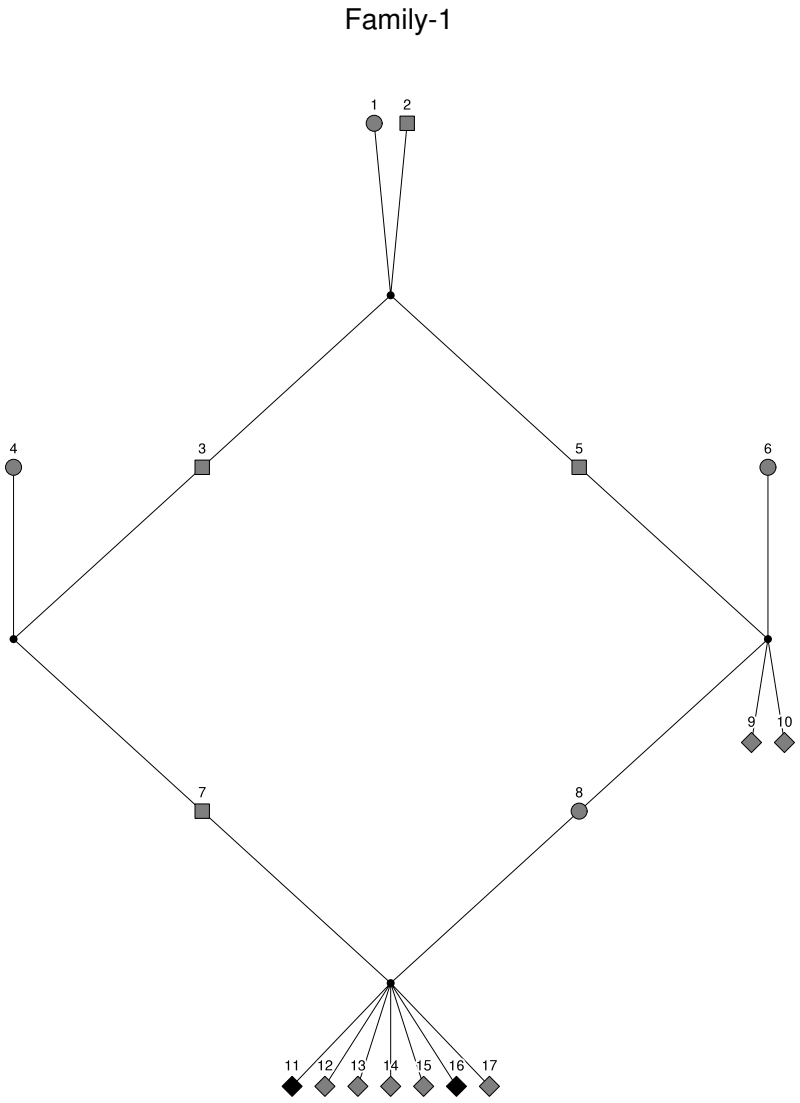

Figure 5-2: Familial hemophagocylic lymphohistiocytosis family 1

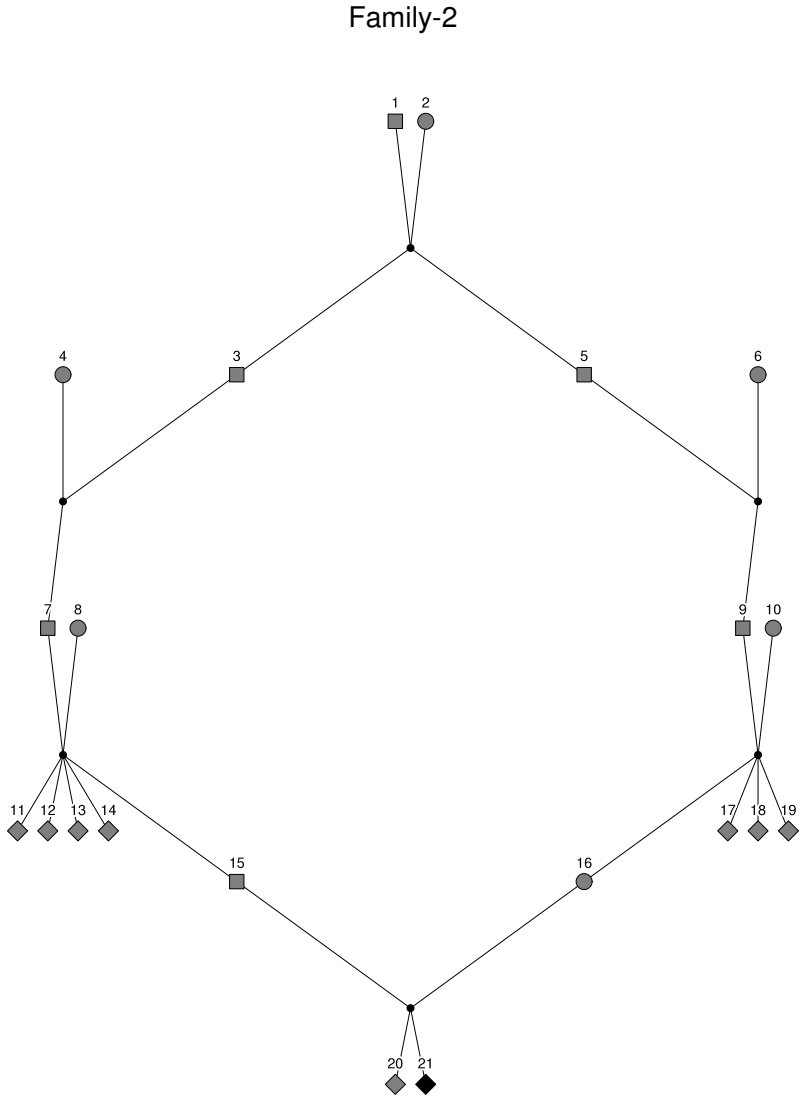

Figure 5-3: Familial hemophagocylic lymphohistiocytosis family 2

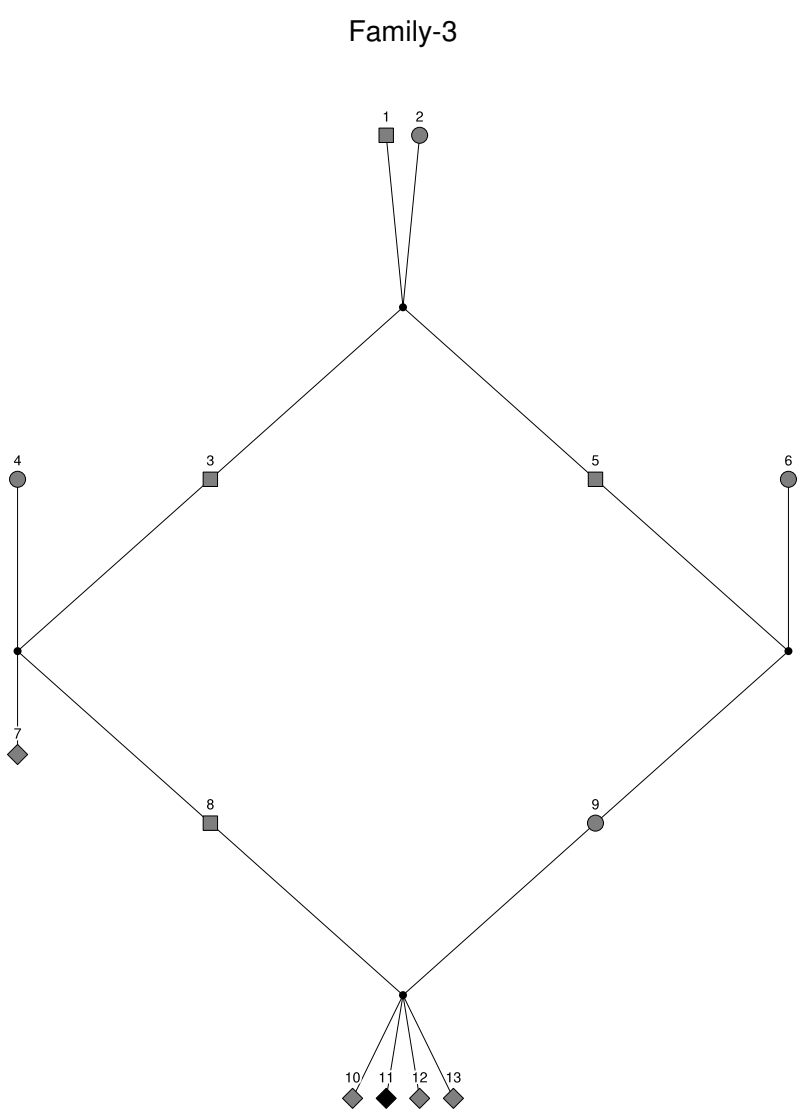

Figure 5-4: Familial hemophagocylic lymphohistiocytosis family 3

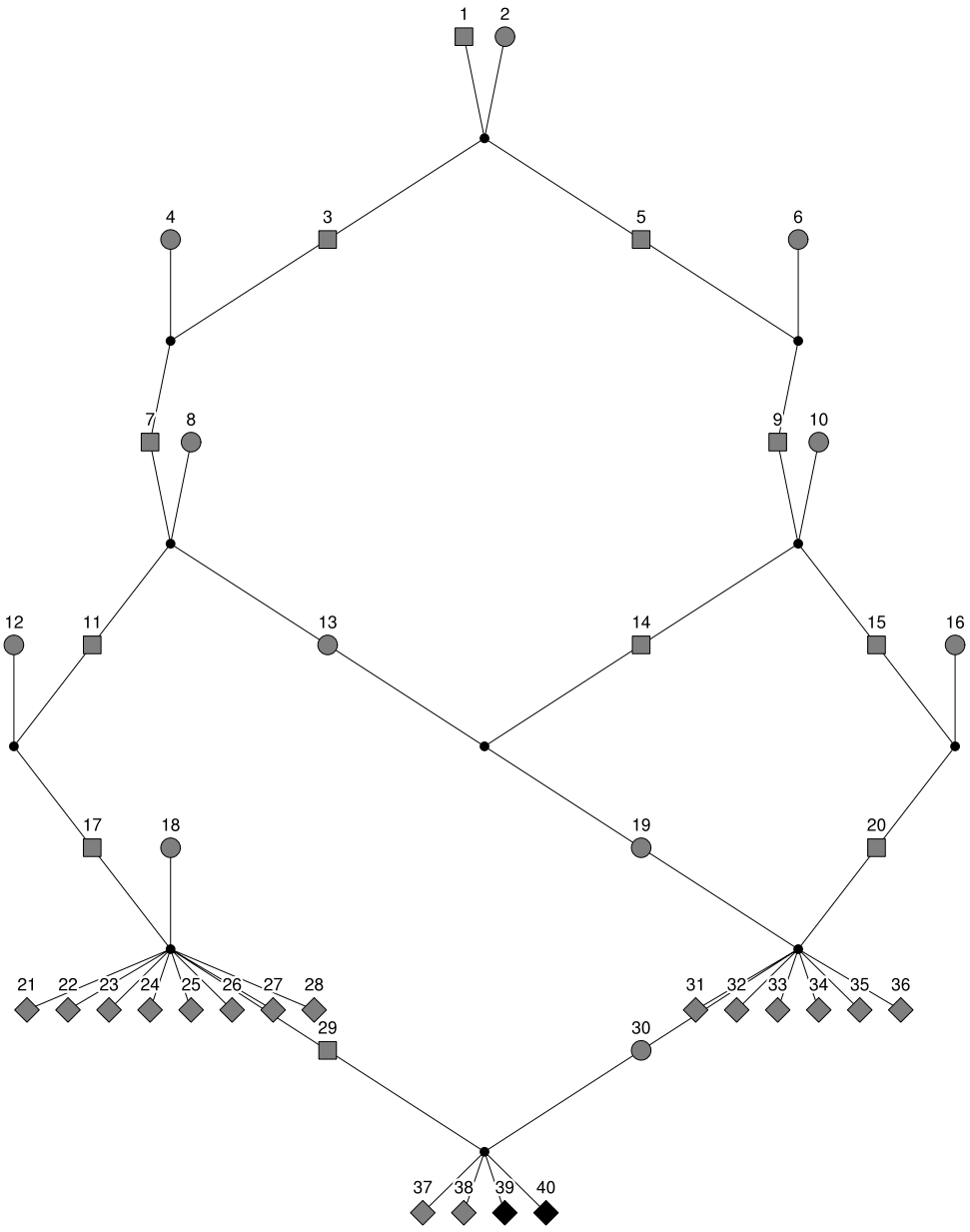

Figure 5-5: Familial hemophagocylic lymphohistiocytosis family 4

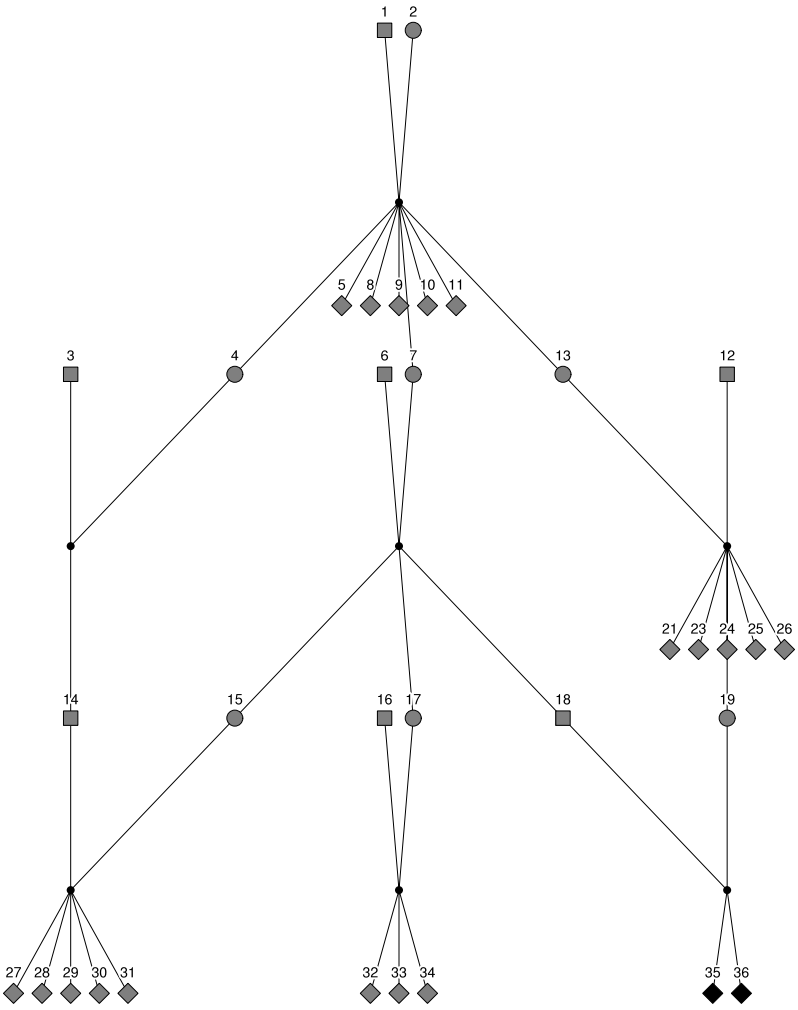

Figure 5-6: Familial hemophagocylic lymphohistiocytosis family 5

and the middle locus is treated as disease locus and used for phenotype matching).

The simulated data is experimented with **LINKAGE**, but owing to presence of many loops, it is quite slow. Because of the loops, the current **VITESSE** could not be used. **GENEHUNTER** is used and individuals that are less informative for linkage are removed from the analysis. The lod scores for the first four pedigrees over 100 replicates are 117.1, 78.6, 91.8, 95.3 for 5cM. As it is very slow, some results are estimated rather than obtained from real simulation and computation, for example, pedigree 5 is assumed to have lod score at least that of pedigree 2. Under 10cM, lod score for pedigree 1 is 92.7, so roughly the lod scores from 5cM to 10cM would be scaled down by a factor of 0.8. Then ELODs for 5cM and 10cM are then 4.614 and 3.691, respectively. This corresponds to χ^2 of 17 for 10cM and a noncentrality parameter of 16. Given that we obtained the lod score distribution under linkage, a lod score of 3 corresponds to χ^2 of 13.8 and a power of 61%, compared to a lod score of 2 with of power 83% (These can be verified via Stata as 1-nchi2(1,16,13.8) and 1-nchi(1,16,9.2)).

The simulation guarantees the proceeding of genotyping, and the linkage of familial homophagocytic lymphohistiocytosis to 9q21.3-22 has been reported in Ohadi et al. (1999): " *Linkage of the disease gene to a region between markers D9S1867 and D9S1790 at 9q21.3-22 is identified by homozygosity mapping in four inbred FHL families of Pakistani descent with a combined maximum multipoint LOD score of 6.05. This is the first genetic locus to be described in FHL. However, homozygosity by descent across this interval could not be demonstrated in an additional affected kindred of Arab origin, whose maximum multipoint LOD score was 20.12. The combined sample revealed significant evidence for linkage to 9q markers (LOD score with heterogeneity, 5.00)*".

The second example is a leukemia family (Figure 5-7, Dr Layton, personal communication). A Mendelian recessive model is also assumed, the gene frequency is about 0.001. Neither **SIMULATE** nor **SLINK** is fast. With 10cM distance between the flanking marker loci the lod score is approximately 1 using 20 replicates of **SLINK**. Linkage analysis is not pursed owing to this result.

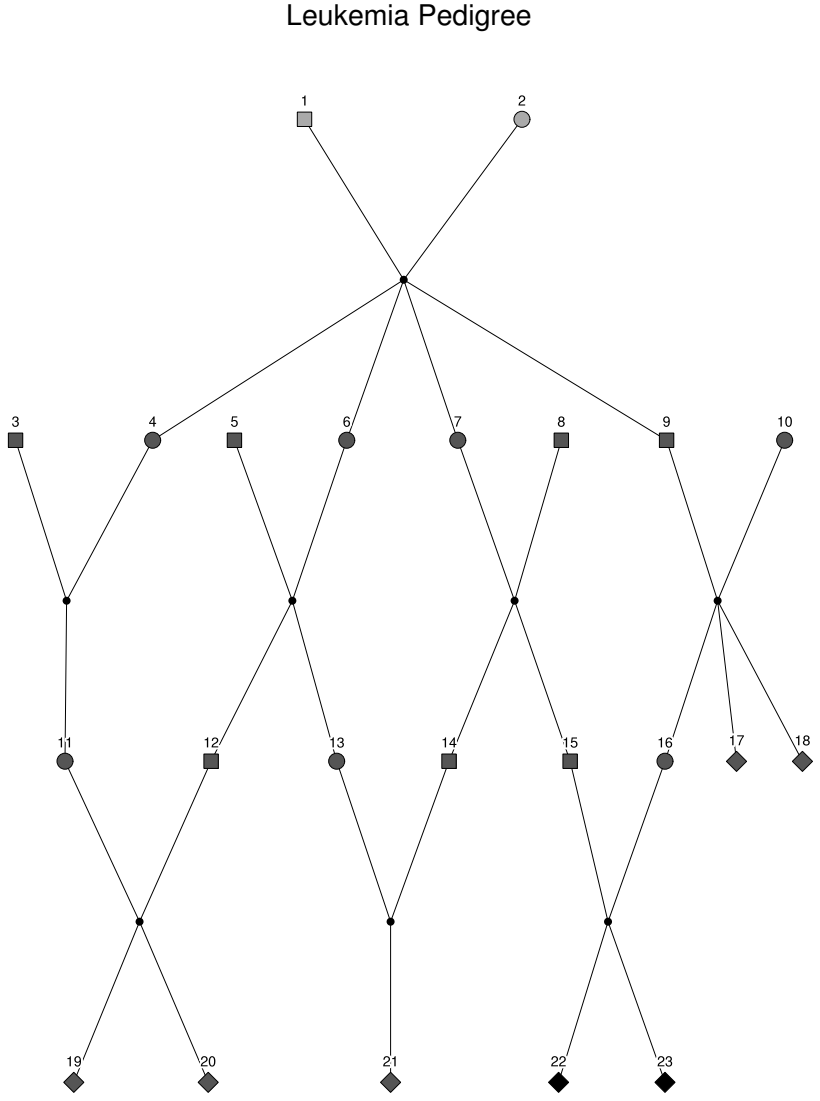

Figure 5-7: The Leukemia family

5.3 Discussion

Simulation of oligogenic traits can be seen as a two-step procedure. First, a microsimulation or demographic simulation is used to get the basic study units, typically either nuclear or extended families. Second, chromosome-based marker loci are generated. Characterisation of these markers is desirable in order to represent features of the human genome and will be served as basis of marker selection. Several programs have been developed for GAWs (Genetic Analysis Workshop) 9 and 10 (e.g. **POPGEN**, **MARKCHAR**, **CHRSIM**). Unlike previous GAWs, GAW 11 started to consider disequilibrium information in the simulated data. When families are involved, this information is mostly among parents, which can be estimated using standard EM algorithm when phase information is unavailable. In fact there are many mechanisms of association so the actual simulation may appear in different forms. Other aspects may be featured. For instance, the RFA for GAW12 described a metabolic model with gene effects operating at different stages and manifest clinical outcomes, and may be highly suggestive of survival analysis and involvement of other particular groups. The events are rare and thus needs stringent ascertainment. It has been shown that chromatid interference does exist and a proper account of crossover interference in gene mapping methods may increase their efficiency. It is therefore desirable to have it incorporated in actual data simulation. Simulation is as useful for validity-checking as for power analysis of test statistics. On both occasions, multiple replicates of a given sample are generated assuming different model parameters and appropriate statistics are obtained. A general discussion, including a population simulation program, **POPSIM** was given by Hampe et al. (1998).

Homozygosity mapping assumes that a rare recessive disorder running in consanguineous families is due to homozygosity by descent (HBD). While powerful, it poses a heavy computational problem, so the analysis is most likely to be good instance for MCMC. Currently no such a program is available.

A brief comparison of **SLINK** with **SIMLINK** is useful. For a given problem **SIMLINK** obtains conditional probabilities of trait genotypes for each pedigree member, a replicate of each of the user-supplied pedigree(s), and lod/location scores for the replicate and their summary statistics. These probabilities are calculated by conditioning on the trait genotypes of their relatives. Simulations are carried out at the specified true recombination

fractions for one marker locus or at the recombination fractions corresponding to the specified map distance for two flanking marker loci. For linked markers, **SIMLINK** obtains the ELOD/location score assuming homogeneity or heterogeneity, and the probability of a maximum lod/location score greater than specified constants. For unlinked markers, it gives ELOD/expected location score for several test recombination fractions/map distances, and the probability of a lod/location score greater than these constants. These give an estimate of the exclusion region when testing for linkage to an unlinked marker (pair). Estimating the probability of a maximum lod/location score greater than 3 for a true recombination fraction of 0.5 gives the probability α of incorrectly concluding linkage to an unlinked marker (pair). The overall probability of type I error can be conservatively corrected if many (pairs of flanking) markers are considered. Assuming that the linkage calculations for the different (pairs of flanking) markers are independent, the overall probability of making a type I error becomes $1-(1-\alpha)^m$, where m is the number of (pairs of flanking) markers. Data from **SIMLINK** can only be analysed under the simulated model (Terwilliger & Ott 1994). Both **SIMLINK** and **SLINK** can generate data under linkage heterogeneity. However if families generated with heterogeneity are analysed assuming homogeneity, the resulting lod scores are too low and do not correspond to the ELOD under heterogeneity; also, the estimates of θ are biased upwards. When generating data under heterogeneity, the easiest solution for analysis is to focus on ELOD; working with the probability that the maximum lod score exceeds a given threshold requires more complicated manipulations. **SLINK** was based on an algorithm of Ott (1989). Suppose there are N people in a pedigree, $\mathbf{x}=(x_1,x_2,\ldots,x_N)$ is the vector of phenotypes, $\mathbf{g}=(g_1,g_2,\ldots,g_N)$ the vector of multilocus genotypes including phase information, and $P(g_i|\mathbf{x})=P(x_1,\ldots,x_i,g_i,x_{i+1},\ldots,x_N)/P(\mathbf{x})$, where the denominator is the likelihood of the entire set of pedigree data, and the numerator is the likelihood of the pedigree data given that individual i has genotype g_i, then the conditional probability distribution of the genotypes given the phenotypes may be calculated by a series of successive risk calculations $P(\mathbf{g}|\mathbf{x})=P(g_1|\mathbf{x})P(g_2|g_1,\mathbf{x})P(\mathbf{g}|g_1,g_2,\mathbf{x})\ldots$. **SLINK** calculates these probabilities (or risks) of all the possible multilocus genotypes with phase $P(g_1|\mathbf{x})$ and, based on these, randomly assigns one of the genotypes to person 1. Then $P(g_2|g_1,\mathbf{x})$ is calculated, taking into account the genotype g_1 just generated, and randomly assign a genotype to person 2, and so on. The process of successive risk calculations continues until all individuals in the pedigree have been assigned a multilocus genotype. This

approach is efficient since once a multilocus genotype has been assigned to one individual, it is known in all subsequent steps. This algorithm permits simulation conditioning on any combination of phenotypic data. For example, if the pedigree were partially typed at a marker of interest, one could simulate conditional on both the disease phenotypes and the marker data currently available. **SLINK** allows one to carry out power calculations by simulating genotypes at one locus given the phenotypes at another locus linked with the first one. There are also faster and parallel versions of **SLINK** called **FASTSLINK** and **PSLINK**.

5.4 Bibliographic notes

GASP uses a similar method for family data generation but does not consider disequilibrium. There have been considerable interest in association study using theory of population genetics, where simulation has been common practice, for example, Kruglyak (1999), Collins et al. (1999). A simulation method for data analysis has been elegantly implemented in **MORGAN**, which was designed for genetic analysis of both qualitative and quantitative traits and has recently incorporated module **Genedrop** for data simulation. It features MCMC in a number of computer programs: **PolyEM**, multiple trait polygenic model with additive effects, **MIXED**, mixed model (major Mendelian genes plus polygenic additive effects), **Multipnt**, Mendelian quantitative trait model with Mendelian markers, **PBGibbs**, Mendelian qualitative trait model, **Autozyg**, multiple Mendelian loci (using meiosis indicators) **Checkers**, pedigree checking, **Loki**, MCMC linkage analysis. A useful expression for constraints over three-locus disequilibria was given by Ayre & Balding (2001).

While actual mechanism for a particular problem may be quite specific, a general program has many factors to take into account. One example is family-size distribution. Cavalli-Sforza & Bodmer (1971) examined progeny size as a measure of fertility, one of the major components of biological fitness. It is in turn measured, crudely, by the mean progeny size after the reproductive period has finished. If all women had the same constant chance of bearing children over a given fixed time period, births would be distributed randomly throughout this period and the distribution of their number would be Poisson. It can be shown that the negative binomial distribution is equivalent to a particular type of mixture of Poisson distributions with different means. More exactly, the means must be distributed

as gamma distribution. In their example, a survey of 307 women, the distribution of observed progeny sizes followed negative binomial distribution given by the terms of $(q-p)^{-n}$, where $q = 1 + p$ and p is positive, $p_r = q^{-n}[(n + r - 1)!]/[r!(n - 1)!](p/q)^r,\ r = 0, 1, \ldots,$ so $p_0 = q^{-n}, p_1 = q^{-n}np/q,$ etc. The mean and variance of this distribution is np and $np(1 + p)$. They found $n = 5.37$ and $p = 0.629$ fit the data well. Ewens (1982) & Morton (1982) showed that family-size distribution should not affect the result of segregation analysis, the distribution of family sizes can affect power calculations because a data set containing 100 two-child families contains less information than, say 100 five-child families. Suarez & Van Eerdewegh (1984) used a geometric sibship size distribution $f(k) = p(1 - p)^{k-2} = 0.4551(1 - 0.4551)^{k-2}, k \geq 2$ so that two siblings occur at less than 50% of the times, three siblings at less than 25% of the time, etc. However, Ewens (1991) gave more justifications why family size distribution is not usually used in a likelihood function involving genetic parameters. Two alternative forms of negative binomial are given here. Suppose that we count the number of Bernoulli trials x required to get a fixed number of successes r then we get negative binomial distributions $P(X = x|r, p) = \binom{x-1}{r-1} p^r(1 - p)^{x-r}$. Since $\{X = x\}$ can only occur when there are $r - 1$ successes in the first $x - 1$ trials and a success on the xth trial which is the product of probability $P(X = x|r, p) = \binom{x-1}{r-1} p^{r-1}(1 - p)^{x-r}$ and p. If we define in terms of the number of failures before the rth success in X trials and $Y = X - r$ and $P(Y = y) = \binom{r+y-1}{y} p^r(1 - p)^y = (-1)^y \binom{-r}{y} p^r(1 - p)^y,\ y = 0, 1, \ldots.$ The mean and variance of this distribution are $\mu = r(1 - p)/p$ and $r(1 - p)/p^2 = \mu + \mu^2/r$. When $r \to \infty, p \to 1, r(1 - p) \to \mu, 0 < \mu < \infty$ then the mean and variance agree with Poisson distribution. If $r = 1$, we have geometric distribution $P(X = x) = p(1 - p)^{x-1}$ with the mean $1/p$ and variance $(1 - p)/p^2$ (Casella & Berger 1990, pp 95-97, 98). Finally, the negative binomial distribution can be obtained as a gamma mixture of Poisson distributions (Knight 2000, p101). The density function of gamma distribution is $f(x) = 1/[\Gamma(\alpha)\beta^\alpha]x^{\alpha-1}e^{-x/\beta},\ \ 0 < x < \infty,\ \ \alpha > 0,\ \ \beta > 0,$ where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}$ is the Gamma function, and α, β are the shape and scale parameters of gamma distribution. This distribution has mean $\alpha\beta$ and variance $\alpha\beta^2$. It is clear that the central χ^2 distribution is just $\Gamma(p/2, 1/2)$, and p is the degree of freedom. Now let $P(x)$ be a Poisson frequency function with mean λ and $f(\lambda)$ be a gamma distribution with mean μ and variance μ^2/α, the mixture distribution has frequency function $g(x) = [\Gamma(x + \alpha)]/[x!\Gamma(\alpha)][\alpha/(\alpha + \mu)]^\alpha[\mu/(\alpha + \mu)]^x$.

Chapter 6

Association in nuclear families

6.1 Introduction

Typical linkage analysis assumes linkage equilibrium between disease locus and markers. Several recent works show that the incorporation of LD information is necessary to increase power (e.g. Terwilliger & Ott 1994; Xiong & Jin 2000). This chapter presents some work on this subject for nuclear families, which is relatively easy to study and generalisable to general pedigrees (e.g. Morton 1955; Elston & Steward 1971; Cannings et al. 1978; Bonney 1986; Martin et al. 2000; Abecasis, Cookson & Cardon 2000).

Assumption of linkage equilibrium between disease locus and markers formed the basis of allele frequency estimation of Boehnke (1991) and maximisation over penetrances of Curtis & Sham (1995). Terwilliger & Ott (1994) suggested that lod score can be more appropriately obtained from the joint likelihood linkage and association, treating haplotype frequencies as nuisance parameters. However as recently pointed out by Xiong & Jin (2000), they did not provide algorithm incorporating LD parameters, nor did they use the correct critical value for the statistics (Self & Liang 1987) and conduct power analysis. Family-based association methods assume a disease locus and a marker locus being either in equilibrium or in disequilibrium. There is a growing consensus that a general likelihood framework combining linkage and association is needed (Thomson 1995a, 1995b; Xiong & Jin 2000; Slager et al. 2001).

Chapter aims

This chapter attempts to use mathematical optimisation in combined linkage and association analysis for nuclear families. These include numerical experiments similar to those in Terwilliger & Ott (1994), Xiong & Jin (2000).

6.2 Methods

Figure 6-1 is from Terwilliger & Ott (1994), as an example of linkage analysis allowing for LD.

Initial haplotype frequencies were obtained from founder individuals using **EH**, and the input data, not given there, is as follows.

```
2 2
3 0 0
1 8 0
0 1 0
```

Note that the first line indicates that there are two markers each with two alleles, the following lines correspond to a 3×3 table formed by the three genotypes at each marker.

Denote the allele frequencies at marker 1 to be p_1, p_2 and those at marker 2 to be q_1, q_2. A single disequilibrium parameter (δ) is sufficient to describe disequilibrium between two biallelic loci. The variation between $\delta = 0$ and $\delta = -p_1 q_1$ or $p_2 q_2$ indicates no and complete linkage disequilibrium. This motivates us to incorporate this parameter into the standard linkage analysis, so that the likelihood calculation is based directly on haplotype frequencies and a proper optimisation procedure is used.

Consider a generalised single locus with allele frequencies p and q for the normal and disease alleles, respectively. The penetrances of genotypes given two, one and no disease alleles are f_2, f_1 and f_0. Consider also a marker locus with n alleles, their associated haplotype frequencies being $h_{11}, ..., h_{1n}, h_{21}, ..., h_{2n}$, conditional on individual's affection status, the genotypic distribution is then obtained according to equation (3.1). Alzheimer's model (Model 3 below) is used to define genotypic probabilities, through which 22 models are generated (see table 6.1) by a SAS program given in Appendix B.

For example with model 1, the haplotype frequencies are as follows

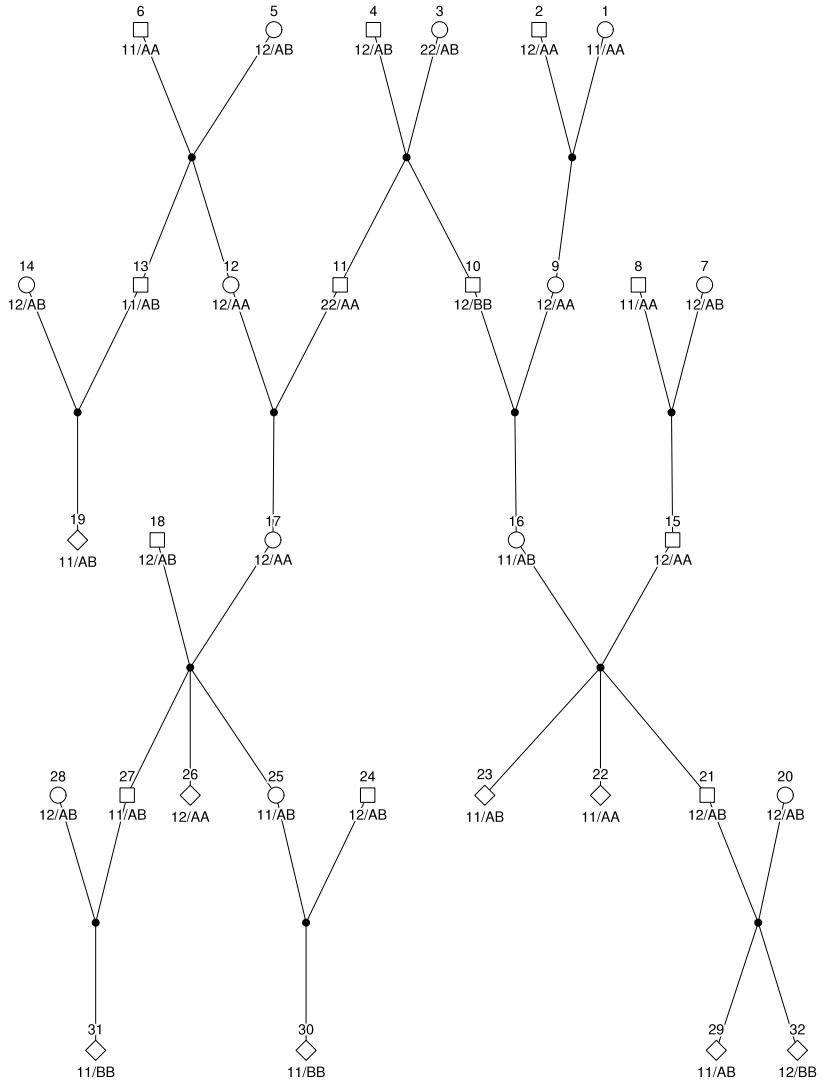

Figure 6-1: Pedigree for association analysis between two marker loci

Table 6.1: The 22 models compatible with Alzheimer's

model	q	δ	f_0	f_1	f_2
1	0.117	0.10027	0.049410	0.21760	0.95834
2	0.130	0.10179	0.045761	0.20521	0.92021
3	0.130	0.11310	0.050000	0.20000	0.80000
4	0.143	0.10365	0.042493	0.19338	0.88008
5	0.156	0.09454	0.034453	0.18342	0.97647
6	0.156	0.10585	0.039598	0.18224	0.83867
7	0.169	0.09707	0.031942	0.17163	0.92221
8	0.182	0.09994	0.029809	0.16087	0.86821
9	0.195	0.09184	0.022419	0.14611	0.95217
10	0.195	0.10315	0.028020	0.15114	0.81520
11	0.208	0.09539	0.021137	0.13677	0.88495
12	0.221	0.08797	0.014451	0.11813	0.96561
13	0.221	0.09928	0.020143	0.12860	0.82105
14	0.234	0.09220	0.013983	0.11129	0.88580
15	0.247	0.08545	0.008373	0.08962	0.95922
16	0.247	0.09676	0.013721	0.10557	0.81225
17	0.260	0.09035	0.008553	0.08621	0.86903
18	0.273	0.08428	0.004232	0.06279	0.93176
19	0.286	0.08986	0.004807	0.06335	0.83481
20	0.299	0.08446	0.001783	0.03972	0.88456
21	0.312	0.07940	0.000161	0.01230	0.93739
22	0.325	0.08599	0.000580	0.02182	0.82145

	marker		
disease	allele 1	allele 2	total
disease allele	0.115479	0.014521	0.13
normal allele	0.001521	0.868479	0.87
total	0.117	0.883	1

Only a subset of these models, here models 1, 3, 11 and 22 are used to experiment on model identification. The second marker is assumed to have three, four, or five equifrequent alleles and in linkage equilibrium with the first (allele frequency 0.1). Nuclear families are generated from the **SIM** program (Chapter 5), with sibship size in accordance with truncated Poisson distribution with parameter 3. Families are kept only when there are two or more family members are affected. For the simulated data, numerical optimisation is conducted to recover the simulated parameters. If q is fixed, it is set to 0.13. If δ is

fixed, it is set to be the maximum 0.1311. -2 log likelihood is minimised over penetrances in addition to different combination of frequency and disequilibrium parameters. The problem is redefined as follows.

When $f_0 = f_1$ and $f_2 = 1$, f_1 achieves the smallest value $[K - q^2]/[1 - q^2]$, $f_1 \geq \max([K - q^2]/[1 - q^2], 0)$, and when $f_0 = 0$ and $f_1 = f_2$ the biggest $K/[1 - (1 - q)^2]$, $f_1 \leq \min(K/[1 - (1 - q)^2], 1)$.

Given f_1, q, if $f_1 < K$, f_2 achieves the smallest value when $f_0 = f_1$, i.e. $[K - (1 - q^2)f_1]/q^2$; If $f_1 > K$, the smallest f_2 is f_1, the largest is when $f_0 = 0$, i.e. $\min([K - 2q(1 - q)f_1]/q^2, 1)$.

Given q, f_1, f_2, $f_0 = [K - 2q(1 - q)f_1 - q^2 f_2]/(1 - q)^2$.

Given the lower bound L and upper bound U of a parameter x, the problem is then converted into a nonlinear optimisation with boundary constraints. $x = L + c(U - L)$, where c varies between 0 and 1.

$$
\begin{aligned}
\min f_1 &= \max((K - q^2)/(1 - q^2), 0) \\
\max f_1 &= \min(K/(1 - (1 - q)^2), 1) \\
f_1 &= \min f_1 + c_1(\max f_1 - \min f_1) \\
\min f_2 &= \begin{cases} f_1, & f_1 > K \\ K - (1 - q^2)f_1)/q^2, & \text{otherwise} \end{cases} \\
\max f_2 &= \min((K - 2qpf_1)/q^2, 1) \\
f_2 &= \min f_2 + c_2(\max f_2 - \min f_2) \\
f_0 &= (K - 2qpf_1 - q^2 f_2)/p^2 \\
\min \delta &= \max(-qp_1, -pp_2) \\
\max \delta &= \min(pp_1, qp_2) \\
\delta &= \min \delta + c_3(\max \delta - \min \delta) \\
K_n &= p^2 f_0 + 2pqf_1 + q^2 f_2
\end{aligned}
$$

Where $c_1 - c_3$ are free parameters for the penetrances and disequilibrium parameter, where K_n is used to check for K. Logistic transformation $f(x) = 1/[1 + \exp(-x)]$ removing the boundary constraints is an alternative but not pursued.

The optimisation is performed with the conjugate gradient method of SAS by repeatedly invoking **LINKMAP**. Statistical power can be appropriately evaluated in terms of sample

sizes based on the log-likelihood ratio statistics. For type I and type II error rates to be $\alpha = 0.0001$, $\beta = 0.10$, the noncentrality parameter is 29.92 for χ^2 degrees of freedom 2.

6.3 Results

Based on Figure 6-1, the likelihood ratio statistic of association versus no association is 12.02, corresponding to a p value of 0.0005. The estimated haplotype frequencies are (0.377219, 0.199704, 0.276627, 0.146450) assuming independence and (0.576921, 0.000002, 0.076925, 0.346152) assuming association; these yield a disequilibrium estimate of 0.199702.

These are used as initial values for **ILINK** to maximise the likelihood over haplotype frequencies using the whole pedigree to establish the phase. When the markers are unlinked the estimate (0.576837, 0.000000, 0.076995, 0.346169) is very similar to those from **EH**. By allowing for linkage between the two markers, and estimate the recombination and haplotype frequencies jointly the results become (0.315367, 0.257574, 0.341849, 0.085210) and disequilibrium value of -0.061852. The values of -2 log-likelihood from **ILINK** are 92.9909 and 97.9948120 so the optimisation process is not satisfactory. Such a seemingly simple example shows that the likelihood surface is not simple.

The same result occurs from SAS optimisation procedure: fixing recombination rate $\theta = 0.5$ and starting with equal haplotype frequencies, -2 log-likelihood value drops from 108.13096 to 92.991507 with estimates (0.5759099, 0, 0.0782231 and 0.345867) under Sun, -2 log-likelihood 85.335481 and haplotype estimates (0.6622288, 0.0000000, 0.1902562, 0.1475151) under DEC Alpha, while allowing for linkage, -2 log-likelihood remains to be 97.975395 with haplotype frequency estimates (0.3168155, 0.2601083, 0.3370297, 0.0860466) under DEC Alpha, suggesting a very flat surface with respect to θ. It seems the failure of SAS may be due to convergence to a local minimum, but log likelihood remains unchanged with moderate change in parameter value.

The result is not examined in detail here for the simulated data from **SIM** due to some vexing problems. First, the scheme from which the sample is simulated may lead to ascertainment bias. Second, it discards families potential for LD, such as family trios (two parents and one child), which is typical data for TDT designs for test of linkage and association. Finally, the implementation in **LINKAGE** is quite limited.

6.4 Discussion

Most current procedures have their limitations. For example, **ILINK** can provide maximum likelihood estimation of allele frequencies but the penetrances are assumed known, and most problems are difficult for the GEMINI routine when highly polymorphic markers are involved. In fact these could be special cases of family likelihood which takes into account allele frequencies, penetrances, disequilibrium, in a combined linkage and association framework. In priciple we can use any linkage analysis routine as function calculator for high-performance optimisation routine, although when the likelihood computation itself is computer intensive, especially for extended pedigrees with many loops.

The likelihood in a typical linkage analysis is affected by founder allele frequencies. It takes little reflection to recognise that founder LD has similar role, and that even families with one affected child can provide useful information about linkage disequilibrium. An explicit likelihood framework is desirable to explore the potential increase in power, although in reality it remains appealing to continue experiments on programs such as **LINKAGE**, with its capability to handle haplotype frequencies.

For single markers it is possible to obtain analytical power via method described in Thomson (1995a, 1995b) or Xiong & Jin (2000). A nuclear family with three siblings (called 3S below in their convention) was examined in Figure 1 of Xiong & Jin (2000). Specifically the family had two unaffected parents and three affected siblings, all were genotyped at a biallelic marker (p_1=frequency of allele 1, p_2=frequency of allele 2) and both parents and the first two siblings were 1/2 heterozygous, while the third sibling is 1/1 homozygous. Now under Mendelian recessive model (p_d=frequency of normal allele, p_D=frequency of disease allele), the likelihood is $4(p_d p_D p_1 p_2)^2[\theta^2(1-\theta)^4 + \theta^5(1-\theta) + (1-\theta)^5\theta + \theta^4(1-\theta)^2]$, with θ being the recombination rate between disease and marker loci. The lod score curve is shown in Figure 6-2.

Assume $p_1 = 0.1$, p_D=0.01, this function is approximately maximised at $\theta = 0.2$ with lod score 0.12. By allowing for LD between disease and marker loci (δ), the lod score function can be expressed as $4(c_1c_4)^2(1-\theta)^4\theta^2 + 4c_1c_2c_3c_4[(1-\theta)^5\theta + \theta^5(1-\theta)] + 4(c_2c_3)^2\theta^4(1-\theta)^2$, where $c_1 = p_1p_D + \delta$, $c_2 = p_1p_d - \delta$, $c_3 = p_2p_D - \delta$, $c_4 = p_2p_d + \delta$. Apparently the lod score 1.06 in their table I would be 0.92. It is rather unfortunate that even in such a simple case it would be difficult with numerical optimisation. We can evaluate the lod score function

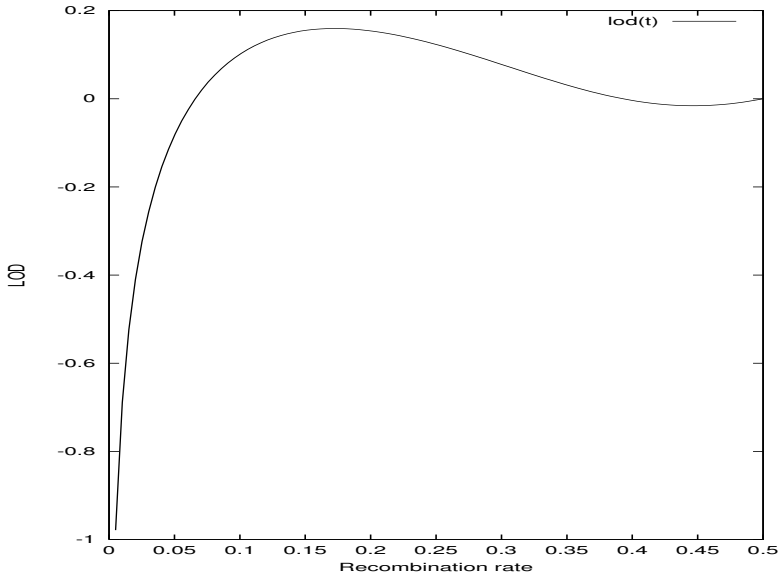

Figure 6-2: LOD score of 3S family assuming linkage equilibrium

using fine grid points of recombination rate and disequibrium value. Many points exceeded their optimal solution (e.g. 1.554 when $\theta = 0.325$, $\delta = 0.009$). Figure 6-3 shows the lod score surface using GNUPlot.

The lod score function indeed favors larger disequilibrium value. In general, let S_{ijkl} be the probability of parental genotype $(i,j) \times (k,l)$) at a multiallelic marker, $S_{ijkl}(\theta, 0) = p_i p_j p_k p_l$ under linkage equilibrium; $S_{ijkl}(\theta, \delta) = p_i p_j p_k p_l W(\theta, \delta)$ under linkage disequilibrium ($W(\theta, \delta)$ is a function of disease/marker allele frequencies and disequilibrium values between them). The noncentrality parameter for family trios (two parents and an affected sibling, called S families) is

$$\lambda_S(\theta, \delta) = 4N \sum_{ijkl} p_i p_j p_k p_l W(\theta, \delta) \ln W(\theta, \delta)$$

For single mutation, and from $\delta(t) = \delta_0 e^{\theta t}$, with t being the number of generations that a single mutation was introduced, they obtained the approximation

$$\lambda_S(\theta, \delta) \approx \frac{4N[\theta^2 + (1-\theta)^2]}{p^2(A)} \left[(f_{DD} - f_{Dd})p_D + (f_{Dd} - f_{dd})p_d\right] \frac{1 - p_1}{p_1} e^{-2\theta t}$$

Where p's and p_D are the marker and disease allele frequencies. $P(A) = \sum_r \sum_t f_{rt} p_r p_t$, $r, t = d, D$. A QBASIC program for this calculation is given in Appendix B. It was surprising

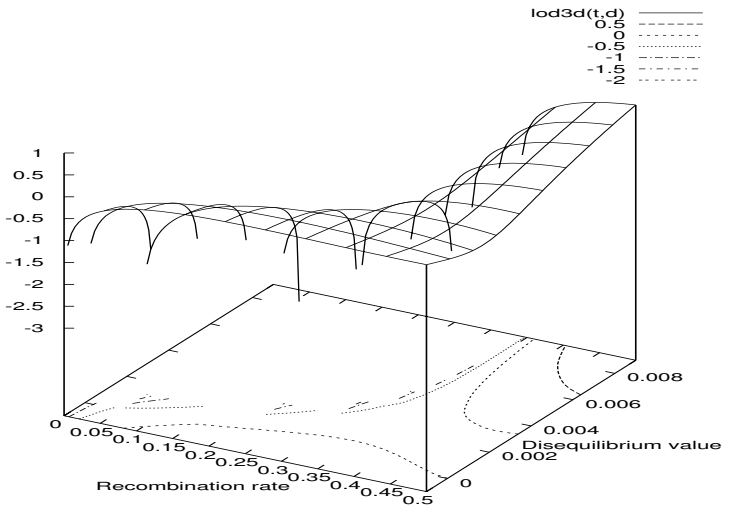

Figure 6-3: LOD score of 3S family assuming linkage disequilibrium

that all the 22 models virtually yielded the same likelihood and noncentrality parameter, which needs to be further investigated.

Lake et al. (2000) also proposed tests of association in the presence of linkage. Under the null hypothesis there is linkage but no association, whereas under the alternaive there are both linkage and association. They also proposed an empirical variance-covariance estimator that is robust to sibling marker-genotype correlation. This is rather appealing but only appropriate for single markers. Recently, Slager et al. (2001), Huang & Jiang (2001) proposed another framework of combined linkage and association analysis. Their method used phase probability as indication of association, for it can be shown that equal phase probabilities for repulsive phases in the usual linkage analysis is true only when there is linkage equilibrium. They expressed the likelihood as a function of both recombination and phase probabilities. Again the threshold for significance is determined by mixture of χ^2 distributions (Self & Liang 1987). They showed that for cases involving two or more affected siblings under a variety of scenarios their method is more powerful than both TDT and affected sibpair mean test.

Mathematical optimisation depends both on the problem and the numerical techniques (e.g. Morton et al. 1983). A successful example is **Mx** (http://views.vcu.edu/mx/), a freely available program for structural equation modelling and widely used for twin data analysis. Both **Mx** and **PAP** use **NPSOL** (http://www.sbsi-sol-optimize.com/NPSOL.htm). Oth-

ers include **CFSQP** (http://gachinese.com/aemdesign/fsqpcontents.htm) and **SEARCH** (http://www.biomath.medsch.ucla.edu/faculty/klange/software.html). To make the models feasible these tools are indispensible.

6.5 Bibliographic notes

The method of Boehnke (1991) is now part of **MENDEL**. There is renewed interest in combined linkage and association analysis (Terwilliger et al. 1992; Terwilliger & Ott 1994; Keavney et al. 1998; Terwilliger & Göring 2000; Xiong & Jin 2000).

GENEHUNTER obtains most probably haplotype frequency estimate via Viterbi algorithm for hidden Markov model. **SimWalk2** (Sobel & Lange 1996) uses simulation-based method to extract haplotype information. **PedManager** (http://waldo.wi.mit.edu /ftp/distribution/software/), **Maga2** (Mukhopadhyay et al. 1999) can also be used to estimate allele frequencies based on family data, but they do not seem to be based on haplotype analysis. It is likely that programs designed for family data could induce bias owing to ascertainment.

Self & Liang (1987) gave asymptotic distribution of log-likelihood ratio statistic under several nonstandard scenarios. Related work on the subject also includes Chernoff (1954), Davis (1977, 1987), Geyer (1994). Genetic examples include Holmans (1993), Chernoff & Lander (1995), Liang & Rathouz (1999).

Chapter 7

Transmission/disequilibrium tests

7.1 Introduction

TDT has become a strong alternative to traditional linkage design, owing particularly to the work of Risch & Merikangas (1996). The power of TDT using both family trios and sibling controls has been investigated by Kaplan et al. (1997), Baur & Knapp (1997) Knapp (1999a, 1999b) Cervino & Hill (2000), among others.

The work of Risch & Merikangas (1996) has provoked heated debates and studies concerning the use of linkage versus association design (Scott et al. 1997; Müller-Myhsok & Abel 1997; Long et al. 1997; Risch & Teng 1998; Morton & Collins 1998; Teng & Risch 1999; Long & Langley 1999; Camp 1997; Knapp 1999a). These are particularly relevant in the context of increasingly dense genetic map. When serving as a guideline for designing a TDT study, it is unfortunate that the paper of Risch & Merikangas (1996) contained a programming error. A correction will shed light on later work such as Camp (1997, 1999) and represent the original conclusions in terms of sibling instead of genotypic relative risks. Comparison with that of a case-control design will further clarify important issues made by those discussants.

Chapter aims

As a practical implementation of Risch & Merikangas (1996) while correcting an error in their program, this chapter examines power of several designs: ASP linkage, singleton/sib-pair TDT and case-controls. The result has been used in a personal communication to

Dr Anthony D. Long (UC Irvine). In addition, the relationship between the case-control method and TDT by Mitchell (2000) is sketched.

7.2 Methods

Assume a disease susceptibility locus has two alleles A and a with population frequencies p and $q = 1 - p$, the penetrances to be f_0, f_1 and f_2 (defined as the conditional probabilities of the disease given 0, 1 and 2 copies of the disease allele), and the recombination fraction between the disease locus and a marker θ. We have population prevalence of the disease $K = p^2 f^2 + 2pqf_1 + q^2 f_0$, additive and dominance variances $V_A = 2pq[p(f_2 - f_1) + q(f1 - f_0)]^2$, $V_D = p^2q^2(f_2 - 2f_1 + f_0)^2$, $\Psi = \theta^2 + (1-\theta)^2$. Suarez et al. (1978) derived IBD probabilities for affected sib-pair as follows

$$\begin{aligned} P(IBD = 0) &= \frac{1}{4} - \frac{(\Psi - .5)V_A + (2\Psi - \Psi^2 - .75)V_D}{4(K^2 + .5V_A + .25V_D)} \\ P(IBD = 1) &= \frac{1}{2} - \frac{2(\Psi^2 - \Psi + .25)V_D}{4(K^2 + .5V_A + .25V_D)} \\ P(IBD = 2) &= \frac{1}{4} + \frac{(\Psi - .5)V_A + (\Psi^2 - .25)V_D}{4(K^2 + .5V_A + .25V_D)} \end{aligned}$$

Under no linkage, the probabilities of affected sib pair sharing 0, 1, and 2 alleles IBD are 1/4, 1/2, and 1/4. Under linkage these sharing probabilities will be different.

Now assume the genotypic relative risks for genotypes AA, Aa, aa are γ^2, γ and 1. The population disease prevalence, the additive and dominance variances are therefore $K = p^2\gamma^2 + 2pq\gamma + q^2 = (p\gamma + q)^2$, $V_A = 2pq(\gamma - 1)^2(p\gamma + q)^2$ and $V_D = p^2q^2(\gamma - 1)^4$. The offspring and sibling relative risks $\lambda_O = 1 + 0.5V_A/K^2$ and $\lambda_S = 1 + (0.5V_A + 0.25V_D)/K^2$ can be rewritten as $\lambda_O = 1 + w$ and $\lambda_S = (1 + 0.5w)^2$, where $w = (pq(\gamma - 1)^2)/(p\gamma + q)^2$. Since the probabilities of siblings sharing none or one allele by descent is $z_0 = 0.25/\lambda_S$ and $z_1 = 0.5\lambda_O/\lambda_S$, the nonshared probability is $Y = 1 - 0.5z_1 - z_0 = 1 - 0.25(\lambda_O + 1)/\lambda_S = (1 + w)/(2 + w)$. The probability of a parent of an affected child (Aff Child) being heterozygous (H) is given by $P(H|\text{Aff Child}) = P(H)P(\text{Aff Child}|H)/P(\text{Aff Child})$, which is $2pq(0.5p(\gamma^2 + \gamma) + 0.5q(\gamma + 1))/(p\gamma + q)^2 = pq(\gamma + 1)/(p\gamma + q)$.

Consider a set of N independent identically distributed random variables B_i with mean 0 and variance 1 under the null hypothesis, mean mean μ and variance σ^2 under the alternative hypothesis. Then the statistic $\sum_{i=1}^{N} B_i/N$ has mean 0 and variance 1 under the null but

mean $\sqrt{N}\mu$ and variance σ^2 under the alternative. The sample size N for a given significance level α and power $1-\beta$ can be estimated by $(Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2$.

For affected sib-pair linkage analysis, the allele shared and nonshared from the ith parent is a random variable denoted by $B_i, i = 1, \ldots, N$ and scored 1 and -1. Under the null hypothesis, the shared and nonshared each has probability 0.5 so the mean and variance of B_i are 0 and 1. Under the alternative $\mu = 2Y - 1$ and $\sigma^2 = 4Y(1-Y)$. Assuming sharing of alleles from both parent to be independent, the required sample size for affected sib-pair under $\theta = 0$ and no linkage disequilibrium is $N = (z_\alpha - \sigma z_{1-\beta})^2/2\mu^2$, $Y = (1+w)/(2+w)$, $w = pq(\gamma-1)/(p\gamma+q)^2$ as above. As for singleton and sib-pair TDT with disease locus or a nearby locus in complete disequilibrium, the number of transmissions of allele A is scored from heterozygous parents. For singleton, denote $B_i = 1/\sqrt{h}, h = pq(\gamma+1)/(p\gamma+q)$ if parent is heterozygous and transmits A, $B_i = 0$ if parent is homozygous, $B_i = -1/\sqrt{h}$ if parent is heterozygous and transmits a. Under the null hypothesis the mean and variance of B_i are 0 and 1, whereas under the alternative they are $\sqrt{h}(\gamma-1)/(\gamma+1)$ and $1-h[(\gamma-1)/(\gamma+1)]^2$, respectively. When sib-pairs instead of singletons are used in TDT analysis, the probability of parental heterozygosity becomes $h = pq(\gamma+1)^2/[2(\gamma p+q)^2 + pq(\gamma-1)^2]$, the same formula for sample size calculation can be applied and the required number of families is half the expected number since there are two independent affected sibs.

This can be compared with those of case-control design using a statistic directly testing association between marker and disease (Long et al. 1997). Suppose we have a randomly ascertained population sample, under HWE multiplicative model in which the genotypic relative risk is γ, the frequencies of the three disease genotypes AA, Aa and aa in cases are $\pi\gamma^2$, $2\pi\gamma pq$, and πq^2, respectively, where π is the "baseline" probability that an individual with *aa* genotype being affected. Similarly the three frequencies in controls are $(1-\pi\gamma^2)p^2$, $2(1-\pi\gamma)pq$ and $(1-\pi)q^2$. A unit χ^2 statistic can be constructed using the following 2×2 table:

	affected genotype	nonaffected genotype	
A	$\pi\gamma^2 p^2 + \pi\gamma pq$	$(1-\pi\gamma^2)p^2 + (1-\pi\gamma)pq$	p
a	$\pi\gamma pq + \pi q^2$	$(1-\pi\gamma)pq + (1-\pi)q^2$	q
	$\pi(\gamma p+q)^2$	$1-\pi(\gamma p+q)^2$	1

or equivalently

	affected genotype	nonaffected genotype	
A	$\pi\gamma p(\gamma p+q)$	$p-\pi\gamma p(\gamma p+q)$	p
a	$\pi q(\gamma p+q)$	$q-\pi q(\gamma p+q)$	q
	$\pi(\gamma p+q)^2$	$1-\pi(\gamma p+q)^2$	1

So that the expected unit frequencies (E) are $\pi p(\gamma p+q)^2$, $\pi q(\gamma p+q)^2$, $p-\pi p(\gamma p+q)^2$ and $q-\pi q(\gamma p+q)^2$, the discrepancies between observed and expected frequencies all have factor $\pi pq(\gamma p+q)(\gamma-1)$ but with negative sign before the second and the third items. The statistic $X^2=\sum(O-E)^2/E$ is then easy to obtain. Under the null hypothesis $\gamma=1$, X^2 has central χ^2 distribution, whereas under the alternative $\gamma>1$ and X^2 follows noncentral $\chi^2_{1,\delta}$ distribution with noncentrality parameter $\delta=[\pi pq(\gamma-1)^2]/[1-\pi(\gamma p+q)^2]$ or $\delta=[\gamma^2 p+q-(\gamma p+q)^2]/[1-\pi(\gamma p+q)^2]$ It is equivalently to derive the power by $Y=\sqrt{X^2}\sim N(\sqrt{\delta},1)$. $1-\beta=\Phi(-Z<Y<Z)$, where Z is a preassigned standard normal deviate.

Let the genome-wide significance level and type II error rate be $\alpha=5\times10^{-8}$ and $\beta=0.2$, respectively, the power is calculated for random ascertainment with three different disease prevalences.

A C program written to implement the above procedure is listed in Appendix B.

7.3 Results

The results are shown in Table 7.1. Column N_L corrects the calculation from the original paper (as $N_{asp}(\times)$). The Alzheimer's disease model is based on Scott et al. (1997).

It turns out that with $\gamma\leq 2$, the expected marker-sharing only marginally exceeds 50% for any allele frequency (p). The use of linkage would need practically nonachievable sample size. Nonetheless, direct tests of association with a disease locus itself can still be quite strong. But it may involve large amount of statistical testing of associated alleles.

The result of the case-control design is shown in Table 7.2. Long et al. (1997) used approximation $1-\beta=\Phi(Z-\sqrt{\delta})$ by taking the area under normal curve in the lower tail as negligible. It seems this approximation is quite good (Columns 3-5).

Clearly it is most favourable for diseases that are relatively common, which has important implications for complex traits. When the disease is relatively common, the disease-allele frequency is intermediate and its effect small, statistical power comparable to that

Table 7.1: Comparison of linkage and association in nuclear families required for identification of disease gene

		Linkage			Association				
γ	p	Y	N_L	P_A	Het_1	N_{tdt}	Het_2	$N_{asp/tdt}$	$N_{asp}(\times)$
4.0	0.01	0.520	6400	0.800	0.048	1098	0.112	235	4260
	0.10	0.597	276	0.800	0.346	150	0.537	48	185
	0.50	0.576	445	0.800	0.500	103	0.424	61	297
	0.80	0.529	3022	0.800	0.235	222	0.163	161	2013
2.0	0.01	0.502	445835	0.667	0.029	5823	0.043	1970	296710
	0.10	0.518	8085	0.667	0.245	695	0.323	264	5382
	0.50	0.526	3751	0.667	0.500	340	0.474	180	2498
	0.80	0.512	17904	0.667	0.267	640	0.217	394	11917
1.5	0.01	0.501	6943229	0.600	0.025	19320	0.031	7776	4620807
	0.10	0.505	101898	0.600	0.214	2218	0.253	941	67816
	0.50	0.510	27040	0.600	0.500	949	0.490	484	17997
	0.80	0.505	101898	0.600	0.286	1663	0.253	941	67816
Alzheimer's:									
4.5	0.15	0.626	163	0.818	0.460	100	0.621	36	109

γ=genotypic risk ratio; p=frequency of disease allele A; Y=probability of allele sharing; N_L=number of ASP families required for linkage; P_A=probability of transmitting disease allele A; Het_1,Het_2=proportions of heterozygous parents; N_{tdt}=number of family trios; $N_{asp/tdt}$=number of ASP families

of standard family-based linkage studies is achieved with a smaller number of randomly sampled individuals. While statistical power in terms of required sample size is important, practicality also needs to be considered. For efficiently diagnosed late onset disease such as non-insulin-dependent diabetes and hypertension, it may not be possible to type parents for affected sibling studies. A further point is that the actual number of individuals genotyped needed would be doubled for linkage, tripled for singleton, and quadrupled for sib-pair, assuming both parents are genotyped in a affected offspring study (Long et al. 1997, also from http://dimitri.ucdavis.edu/association_study/index.html). Long et al. (1997) noted when there is actual marker-disease data, a Fisher's exact test can be used in the case of population sample to detect association.

Table 7.2: Estimated sample sizes required for association detection

		Long et al. (1997)			Actual calculation		
γ	p	1%	5%	10%	1%	5%	10%
4.0	0.01	46681	8959	4244	46637	8951	4240
	0.10	8180	1570	744	8172	1568	743
	0.50	10890	2090	990	10880	2088	989
	0.80	31473	6040	2861	31444	6035	2859
2.0	0.01	403970	77530	36725	403593	77457	36690
	0.10	52709	10116	4792	52660	10106	4787
	0.50	35284	6772	3208	35252	6765	3205
	0.80	79390	15236	7217	79316	15222	7211
1.5	0.01	$> 10^6$	307055	145447	1598429	306769	145312
	0.10	192104	36869	17464	191925	36834	17448
	0.50	98012	18810	8910	97921	18793	8902
	0.80	192104	36869	17464	191926	36834	17448

7.4 Discussion

Generalisations of Risch & Merikangas (1996) have been considered by Camp (1997, 1999) and Knapp (1999a). One criticism was that genotypic relative risk γ rather than sibling relative risk λ_s was employed–since even with large γ, λ_s could still be small, this was remedied in Risch (1997). Moreover, one should not rely too much on complete disequilibrium assumption (Müller-Myhsok & Abel 1997). With the same genetic model and one biallelic marker with alleles B and b of frequencies m and $1-m$, they considered the following table,

	B	b	
A	$pm+\delta$	$p(1-m)-\delta$	p
a	$(1-p)m-\delta$	$(1-p)(1-m)+\delta$	$1-p$
	m	$1-m$	1

where the linkage disequilibrium parameter $\delta = P(AB) - pm$ achieves its maximum when $P(AB)$ is $\min(m,p)$. Let $\alpha_1 = P(A|B) = (pm+\delta)/m = p + \delta/m$ and $\alpha_2 = P(A|b) = [p(1-m)-\delta]/(1-m) = p - \delta/(1-m)$, for TDT with trios the probability that Bb parent transmits B to the affected child is $\tau(B) = P(\text{aff}|B)/[P(\text{aff}|B) + P(\text{aff}|b)]$, the prior probabilities of transmitting B and b are both 0.5, and $P(\text{aff}|B) = [\gamma\alpha_1 + (1-\alpha_1)]D$ and $P(\text{aff}|b) = [\gamma\alpha_2 + (1-\alpha_2)]D$, where D is the probability that is a subject is affected given he carries allele a. It turns that out $\tau(B) = [1+(\gamma-1)\alpha_1]/[2+(\gamma-1)(\alpha_1+\alpha_2)]$,

which only reduces to $\gamma/(1+\gamma)$ when $m = p$ whereas when m/p departs from unity the power diminishes substantially. Abel & Müller-Myhsok (1998) also expressed TDT as $\Lambda = 2n\ln[q\ln(q) + (1-q)\ln(1-q) - \ln(0.5)]$, n being the number of heterozygous parents, q being the probability Bb transmitting B, and considered the difference between Λ and the classic TDT $d(q) = 2n[q\ln(q) + (1-q)\ln(1-q) - \ln(0.5) - 2(q-0.5)^2]$. Assuming that under the null hypothesis $q \sim N(\tau(B), \sigma^2)$, $\sigma^2 = \tau(B)(1-\tau(B))/n$, the number of heterozygous parents is obtained by solving the following equation

$$2n[q_\beta \ln(q_\beta) + (1-q_\beta)\ln(1-q_\beta) - \ln(0.5)] = Z_\alpha^2$$

where $q_\beta = \tau(B) - \sigma Z_{1-\beta}$. From n, the required number of families $N = n/2h$ can be estimated, h is the probability that a parent with an affected child is heterozygous given by Risch & Merikangas (1996), or by Müller-Myhsok & Abel (1997) as $u/[u + m^2[1 + (\gamma - 1)\alpha_1] + (1-m)^2[1 + (\gamma-1)\alpha_2]$ with $u = m(1-m)[2 + (\gamma-1)(\alpha_1 + \alpha_2)]$. Setting $\gamma = 4$, $\alpha = 5 \times 10^{-8}$ and $1 - \beta = 0.80$, for the maximum likelihood test of $p = 0.5$ (ML-TDT), $N = 139$ for $p = m = 0.10$ and $N = 96$ for $p = m = 0.50$, compared with 150 and 103 in Table 7.1.

A thorough power analysis of various tests would be beyond this chapter. However a number of observations can be made. First, TDT is by no means a substitute for other epidemiological, linkage and case-control association designs. Second, the basic principle should apply to both quantitative and qualitative traits. Third, it is not limited to nuclear families. Fourth, multiply linked marker can preferably be incorporated for a haploptype disequilibrium test. Fifth, there is no reason why TDT analysis should be separated from other analyses.

Ideally, unrelated individuals data from a population, as well as family data, with information from multiple markers, can be combined into a unified framework and subject to analyses that are powerful, robust and fast. For example, methods as in **TRANSMIT** is very appealing but sometimes it may be too slow when the Monte Carlo option is used.

7.5 Bibliographic notes

Mitchell (2000) obtained a simple relation between case-control and TDT. In her notation, the magnitude of disease-marker association detected by TDT can be estimated by binomial

proportion $T = M_1/(M_1 + M_2)$, for the number of times high (M_1) and low (M_2) risk alleles are transmitted, respectively. Let m=allele frequencies of high risk marker allele, M_1, g_i=P(genotype i|aff)P(aff)/P(genotype i), i=2,1,0 for M_1M_1, M_1M_2, M_2M_2. This statistic can be rewritten as a function of the relative frequencies of the marker genotypes among the affected offspring of a heterozygous parent as follows,

Child genotype	Allele inherited	P(genotype)	P(aff|genotype)	Frequencies
M_1M_1	M_1	$0.5m$	g_2	$0.5m(g_2)$
M_1M_2	M_1	$0.5(1-m)$	g_1	$0.5(1-m)(g_1)$
M_1M_2	M_2	$0.5m$	g_1	$0.5m(g_1)$
M_2M_2	M_2	$0.5(1-m)$	g_0	$0.5(1-m)(g_0)$

Now let $A = 0.5m(g_2) + 0.5(1-m)(g_1)$, $B = 0.5m(g_1) + 0.5(1-m)(g_0)$, then $T = A/(A+B)$, T could be referred to power/sample size formula for proportion.

Some examples in Mitchell (2000) are as follows,

Disease	Marker	T[case-control, TDT (C.I.)]
Cleft lip/palate	TGFA	0.67 0.77 (0.63-0.91)
Cleft palate	MSX1	0.59 0.68 (0.51-0.85)
Spina bifida	5,10 mthfr	0.60 0.56 (0.49-0.63)
IDDM	5'FP	0.77 0.63 (0.54-0.72)

The estimate of T is outside confidence interval (C.I.) predicted by TDT only in the last example. For the first disorder, T_1=0.62, T_0=0.5, type I error α=0.05, type II error β=0.20, then the required sample size becomes $N = [1.64[0.5(1-0.5)]^{0.5} + 0.84[0.62(1-0.62)]^{0.5}]^2/(0.62-0.50)^2 \approx 105$, or N=266 for α=0.001.

McGinnis (1998, 2000) further clarified the relative power of TDT and affected sib-pair method, including expression of general mode of inheritance. A comprehensive evaluation by computer simulation was given by Kaplan et al. (1997), Monks et al. (1998). Others included Baur & Knapp (1997), Trégouët at al. (2001).

Cervino & Hill (2000) compared **TRANSMIT**, **SIBASSOC/STDT** and **RCTDT** (Knapp 1999b) in a variety of scenarios. When one or two parents are missing, the presence

of population substructure. From their simulation, the classic "likelihood-ratio association test" would result in high type I error when there is a population substructure suggesting simple parameterisation in terms of transmission and different allele frequencies may not be enough to characterise the effect of substructure whereas introducing more parameters would increase the degrees of freedom and reduce power. It seems **TRANSMIT** is quite robust to substructure and provides both correct type I error and reasonable power. Since **TRANSMIT** is based on a multiplicative model it is also most powerful when this is true. **RCTDT** is very appealing since it combines both vertical and horinzontal approaches, i.e., allowing for analysis of incomplete families by reconstructing the missing parents and by comparing affected and unaffected siblings inside families.

Both Abel & Müller-Myhsok (1998) and Weinberg et al. (1998) noted that likelihood ratio test can be more powerful than TDT. In general, statistical power of TDT depends on the number of parents heterozygous for particular alleles. The magnitude of transmission disequilibrium of a marker allele depends on factors such as genotype relative risk of the susceptibility locus, the genetic distance between disease and marker loci, and the amount of linkage disequibrium (Schaid 1996). Schaid (1998), Schaid & Rowland (1998), Schaid (1999a, 1999b), Schaid & Rowland (2000), Knapp (1999a, 1999b, 1999c), Horvath et al. (2000) gave more accounts about TDT. A more recent summary of the statistical issues in TDT was given by Zhao (2000).

An important remark is that case-control design is commonly believed to be more powerful assuming there is no population stratification (Morton & Collins 1998; Devlin & Roeder 1999; Bacanu et al. 2000; Risch 2000; Pritchard & Rosenberg 1999; Pritchard et al. 2000; Reich & Goldstein 2001). A useful development was made by Seltman et al. (2001) which proposed to associate the evolutionary tree with TDT. It would be of interests to use the same set of haplotype frequencies in the Fragile X example to investigate the power of TDT (Sham & Curtis 1995a). A SAS program is available from the section website.

Chapter 8

Bells and Whistles

This chapter gives a short summary of the major findings and discusses possible extensions.

8.1 Major findings

Several scenarios of statistical power analysis in human genetic linkage and association have been investigated. The use of computer simulation methods and power of TDT and other association designs are also explored. It is found that assuming full marker informativeness, parametric and nonparametric linkage statistics have almost equal power, but nonparametric linkage statistics are anticonservative in some families. In case-control association analysis with unrelated individuals, a heterogeneity statistic has comparable power to parametric tests and comparable to the ordinary likelihood ratio test for contingency table. Depending on mode of inheritance, the classic case-control and family designs are equally important. With availability of large number of SNPs, the role of other types of markers in detecting LD can not be overlooked. The power of different designs may be variable depending on the underlying disease model.

The investigation of statistical power has important implications. It should be seen as part of a dynamic process in line with methodological development. Indeed genetic analysis has been one of the most fascinating and motivating fields for modern probabilistic and statistical theory, and an eminent recent feature is that the sheer amount of information will surpass any sophisticated mathematical model and computing resource available. As for human genetic linkage and association analysis, the large number of genetic markers in junction with the complexity of large family poses special challenges, and methods for

combined linkage and association analysis for multiplex families and other data sources such as unrelated individuals are yet to be fully integrated.

8.2 Possible extensions

While many other aspects such as QTL analysis and DNA pooling are left out in this thesis, possible extensions, by chapter, are sketched as follows.

- *Linkage tests for arbitrary pedigrees.* For large pedigrees, complete enumeration and full marker informativeness will soon become unwieldy, so simulation-based methods must be adopted. Other factors such as marker allele frequencies will have to be considered.

- *Case-control analysis.* The major problem is the large number of parameters when multiple and/or multiallelic markers are involved, so heuristic methods and parsimonious models have to be used.

- *Multilocus model for fine mapping.* The deterministic model can to be strengthened with more loci. It would be more desirable to take into more account of human evolutionary history, i.e., by stochastic simulations and coalescent models.

- *Conditional simulation for homozygosity mapping.* While there are programs for MCMC linkage analysis, programs for routine power analysis remain to be developed.

- *LD mapping using families.* While this thesis only reveals some issues of combined linkage and association, most computer programs are not ready for practical data analysis. Good optimisation techniques and programs are necessary but not properly incorporated. Alternative models could also be investigated. It is desirable to obtain the power table for the generated models along the line of Xiong & Jin (2000).

- *Comprehensive power analysis of TDT method.* This should apply to multiplex pedigrees with multiple markers, and possibly with other factors such as incomplete information, missing data, covariates and human evolutionary history.

Complex traits need a good working model to distinguish the contributing factors such as small sample size, insufficient markers, multiple testing, nonuniform linkage disequilibrium

across genome, genotype error, and phenotype definition (Kruglyak & Lander 1995, Risch 2000; Göring & Terwilliger 2000; Cardon & Bell 2001), heterogeneity, gene-environmental interactions, population history not fully used, etc. The demographic and evolutionary history of the current world population, while important for applications such as admixture mapping (e.g. Laan & Pääbo 1997; Chakraborty & Weiss 1988), remains uncertain, and genetic distance does not necessarily mean physical distance (Morton 2000). LD is not merely dependent upon physical distance, but also on differing allele frequencies and age of mutation. High-throughput genome screening may need stringent prior probability for linkage and association (Risch 2000).

It has been several years after Lander & Schork (1994) reviewed then the research paradigms in genetic analysis of complex traits. As the Human Genome Project is approaching its targets these have gradually shifted (e.g., Risch & Merikangas 1996; Schork et al. 1998; Guo & Lange 2000; Johnson & Todd 2000). This change is attributable to the interplay of both theory and application. A classic example was Spielman et al. (1993) in re-analysing of their insulin-dependent diabetes mellitus and a 5' flanking polymorphism of the insulin locus, which also initiated interests on TDT. Consortium work such as GAW has been motivated to develop and evaluate various statistical techniques. Methodological advance requires study designs that combine recent developments in statistics, biology and epidemiology.

8.3 Epilogue

This thesis is largely done as a result of ongoing papers I have been involved as a PostDoc research worker in several Wellcome project grants. The consequence is two-fold. On the one hand, it does not fully reflect my work such as haplotype analysis; on the other, given the dynamic nature of the field and my own limitation, it would have been a dream to provide a comprehensive account of all topics in the thesis, and it is more appropriate to see it as a timely summary. Initially trained in medicine and medical statistics, it is a result of many years of *wading into the swamp* of literature of statistical genetics, mathematical models, algorithms, data structure and computer programming. The joy of being involved and presenting work in this exciting area is beyond words.

Appendix A

Gene counting method

A more systematic and comprehensive derivation of simple cases of gene counting than previously described (Terwilliger & Ott 1994; Sham 1998) is given here. This includes marker-marker analysis of two and three biallelic markers and marker-disease analysis of one and two biallelic markers with a putative disease locus. A summary of symbols used is given in table A.1.

Table A.1: Symbols and their meanings

symbol	meaning
p	normal allele (D) frequency of a biallelic disease locus, or frequency of the first allele of a three-allele disease locus
q	disease allele (d) frequency of a biallelic disease locus, or frequency of the second allele of a three-allele disease locus
r	frequency of the third allele of a three-allele disease locus
f	disease penetrances given genotypes at disease locus
s	normal penetrances given genotypes at disease locus
K	disease prevalence in population
Q	normal prevalence in population
a	cases
u	controls
n	cell count(s) of observed contingency table
N	total number of individuals
c	haplotype count
h	haplotype frequency
g	genotypic frequency
D	LD parameter

Two biallelic markers

Let alleles at each marker be 1 and 2, the haplotypes they form be 11, 12, 21, 22 and frequencies h_{11}, h_{12}, h_{21}, h_{22}. The observed data can be organised into table A.2.

Table A.2: Genotype counts for biallelic markers

	marker 2		
marker 1	1/1	1/2	2/2
1/1	n_0	n_1	n_2
1/2	n_3	n_4	n_5
2/2	n_6	n_7	n_8

Note cells of this table are numbered using the convention of C programming language. The haplotype contribution from each cell is unambiguous except for n_4, where EM algorithm comes in. The doubly heterozygous cell n_4 can be from two possibilities or phase, 21/12 or 22/11 haplotypes. The E-step obtains the expected probabilities $\alpha^4 = h_{21}h_{12}/(h_{21}h_{12} + h_{22}h_{11})$ for 21/12, $1 - \alpha^4$ for 22/11, and haplotype counts $c_{11} = 2n_0 + n_1 + n_3 + (1 - \alpha^4)n_4$ $c_{12} = n_1 + 2n_2 + n_5 + \alpha^4 n_4$, $c_{21} = n_3 + 2n_6 + n_7 + \alpha^4 n_4$, $c_{22} = n_5 + n_7 + 2n_8 + (1 - \alpha^4)n_4$. Note that the sequence number of a cell appears in the superscript for any expected probability, and that phase number appear as subscript when there are more than two phases. Now the M-step is to obtain haplotype frequencies which are simply $c_{ij}/(2N)$, $i, j = 1, 2$, $N = \sum_{i=0}^{8} n_i$. Start with independent haplotype frequencies (being product of the constituent allele frequencies), the haplotype frequencies are updated iteratively. The genotypic probabilities g, $i = 0, \ldots, 8$, are then obtained as table A.3.

Table A.3: Genotypic probabilities for two biallelic markers

	marker 2		
marker 1	1/1	1/2	2/2
1/1	h_{11}^2	$2h_{11}h_{12}$	h_{12}^2
1/2	$2h_{21}h_{11}$	$2(h_{21}h_{12} + h_{22}h_{11})$	$2h_{22}h_{12}$
2/2	h_{21}^2	$2h_{21}h_{22}$	h_{22}^2

The log-likelihood is based on multinomial distribution $l = \sum_{i=0}^{8} n_i \ln(g_i)$. Denote the log-likelihood assuming linkage equilibrium as l_0, that from the gene counting procedure as l_1, $2(l_1 - l_0)$ has asymptotic χ^2 distribution with $(2 \times 2 - 1) - [(2 - 1) + (2 - 1)] = 1$ degree of freedom.

Three biallelic markers

The observed contingency table is now as table A.4.

Table A.4: Genotype count for three biallelic markers

		marker 3		
marker 1	marker 2	1/1	1/2	2/2
	1/1	n_0	n_1	n_2
1/1	1/2	n_3	n_4	n_5
	2/2	n_6	n_7	n_8
	1/1	n_9	n_{10}	n_{11}
1/2	1/2	n_{12}	n_{13}	n_{14}
	2/2	n_{15}	n_{16}	n_{17}
	1/1	n_{18}	n_{19}	n_{20}
2/2	1/2	n_{21}	n_{22}	n_{23}
	2/2	n_{24}	n_{25}	n_{26}

With slight complication we obtain our haplotype counts as follows

$$
\begin{aligned}
c_{111} &= 2n_0 + n_1 + n_3 + \alpha^4 n_4 + n_9 + (1-\alpha^{10})n_{10} + (1-\alpha^{12})n_{12} + \alpha_4^{13} n_{13} \\
c_{112} &= n_1 + 2n_2 + n_5 + (1-\alpha^4)n_4 + \alpha^{10} n_{10} + n_{11} + \alpha_3^{13} n_{13} + (1-\alpha^{14})n_{14} \\
c_{121} &= n_3 + 2n_6 + n_7 + (1-\alpha^4)n_4 + \alpha^{12} n_{12} + \alpha_2^{13} n_{13} + n_{15} + (1-\alpha^{16})n_{16} \\
c_{122} &= n_5 + n_7 + 2n_8 + \alpha^4 n_4 + \alpha_1^{13} n_{13} + \alpha^{14} n_{14} + \alpha^{16} n_{16} + n_{17} \\
c_{211} &= n_9 + \alpha^{10} n_{10} + \alpha^{12} n_{12} + \alpha_1^{13} n_{13} + 2n_{18} + n_{19} + n_{21} + \alpha^{22} n_{22} \\
c_{212} &= (1-\alpha^{10})n_{10} + n_{11} + \alpha_2^{13} n_{13} + \alpha^{14} n_{14} + n_{19} + 2n_{20} + (1-\alpha^{22})n_{22} + n_{23} \\
c_{221} &= (1-\alpha^{12})n_{12} + \alpha_3^{13} n_{13} + n_{15} + \alpha^{16} n_{16} + n_{21} + (1-\alpha^{22})n_{22} + 2n_{24} + n_{25} \\
c_{222} &= \alpha_4^{13} n_{13} + (1-\alpha^{14})n_{14} + (1-\alpha^{16})n_{16} + n_{17} + \alpha^{22} n_{22} + n_{23} + n_{25} + 2n_{26}
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha^4 &= h_{111}h_{122}/(h_{111}h_{122} + h_{121}h_{112}) \\
\alpha^{10} &= h_{211}h_{112}/(h_{211}h_{112} + h_{212}h_{111}) \\
\alpha^{12} &= h_{211}h_{121}/(h_{211}h_{121} + h_{221}h_{111}) \\
\alpha_1^{13} &= h_{211}h_{122}/(h_{211}h_{122} + h_{212}h_{121} + h_{221}h_{112} + h_{222}h_{111}) \\
\alpha_2^{13} &= h_{212}h_{121}/(h_{211}h_{122} + h_{212}h_{121} + h_{221}h_{112} + h_{222}h_{111}) \\
\alpha_3^{13} &= h_{221}h_{112}/(h_{211}h_{122} + h_{212}h_{121} + h_{221}h_{112} + h_{222}h_{111}) \\
\alpha_4^{13} &= h_{222}h_{111}/(h_{211}h_{122} + h_{212}h_{121} + h_{221}h_{112} + h_{222}h_{111})
\end{aligned}
$$

$$
\begin{aligned}
\alpha^{14} &= h_{212}h_{122}/(h_{212}h_{122}+h_{222}h_{112}) \\
\alpha^{16} &= h_{221}h_{122}/(h_{221}h_{122}+h_{222}h_{121}) \\
\alpha^{22} &= h_{211}h_{122}/(h_{211}h_{222}+h_{221}h_{212})
\end{aligned}
$$

are expected contributations based on haplotype frequency estimates from last iteration. For example for cell 13 there are 4 possible phases, the probability of each phase is given by δ_i, $i = 1, \ldots, 4$. As before $h_{ijk} = c_{ijk}/(2N)$, $i, j, k = 1, 2$, and $N = \sum_{i=0}^{26} n_i$. The genotypic probabilities, g_i, $i = 0, \ldots, 26$ are now in table A.5.

Table A.5: The genotypic probabilities for three biallelic markers

		marker 3		
marker 1	marker 2	1/1	1/2	2/2
	1/1	h_{111}^2	$2h_{112}h_{111}$	h_{112}^2
1/1	1/2	$2h_{111}h_{121}$	$2(h_{112}h_{121}+h_{111}h_{122})$	$2h_{112}h_{122}$
	2/2	h_{121}^2	$2h_{121}h_{122}$	h_{122}^2
	1/1	$2h_{211}h_{111}$	$2(h_{211}h_{112}+h_{212}h_{111})$	$2h_{212}h_{112}$
1/2	1/2	$2h_{211}h_{121}$	$2(h_{211}h_{122}+h_{212}h_{121})$	$2h_{212}h_{122}$
		$+2h_{221}h_{111}$	$+2(h_{221}h_{112}+h_{222}h_{111})$	$+2h_{222}h_{112}$
	2/2	$2h_{221}h_{121}$	$2(h_{221}h_{122}+h_{222}h_{121})$	$2h_{222}h_{122}$
	1/1	h_{211}^2	$2h_{211}h_{212}$	h_{212}^2
2/2	1/2	$2h_{211}h_{221}$	$2(h_{211}h_{222}+h_{221}h_{212})$	$2h_{212}h_{222}$
	2/2	h_{221}^2	$2h_{221}h_{222}$	h_{222}^2

The log-likelihood can be calculated in a similar fashion as for two biallelic markers but now with $(2 \times 2 \times 2 - 1) - [(2-1) + (2-1) + (2-1)] = 4$ degrees of freedom for the log-likelihood ratio χ^2 test.

A disease locus and a marker

Assume the disease locus to be biallelic with disease d and normal alleles D, their frequencies to be q and $p = 1 - q$, the genotypes dd, Dd and DD have disease penetrances f_{dd}, f_{Dd}, f_{DD}, and normal penetrances s_{dd}, s_{Dd}, s_{DD}, so that disease and normal prevalences to be $K = q^2 f_{dd} + 2pqf_{Dd} + p^2 f_{DD}$ and $Q = 1 - K$, respectively. The conditional probabilities of genotypes dd, Dd, DD, $\frac{q^2 f_{dd}}{K}$, $\frac{2pqf_{Dd}}{K}$, $\frac{p^2 f_{DD}}{K}$ for being a case, and $\frac{q^2 s_{dd}}{Q}$, $\frac{2pqs_{Dd}}{Q}$, $\frac{p^2 s_{DD}}{Q}$ for being a control suggests Table A.6. Each cell of the contingency table contains two component, one from cases and one from controls.

We can use table A.6 to obtain haplotype frequencies as before, $\alpha^4 = h_{D1}h_{d2}/(h_{D1}hd_{d2}+$

Table A.6: Initial count table for disease locus and a marker

disease locus	marker 1/1	marker 1/2	marker 2/2
d/d	$\frac{q^2 f_{dd} a_0}{K} + \frac{q^2 s_{dd} u_0}{Q}$	$\frac{q^2 f_{dd} a_1}{K} + \frac{q^2 s_{dd} u_1}{Q}$	$\frac{q^2 f_{dd} a_2}{K} + \frac{q^2 s_{dd} u_2}{Q}$
D/d	$\frac{2pq f_{Dd} a_0}{K} + \frac{2pq s_{Dd} u_0}{Q}$	$\frac{2pq f_{Dd} a_1}{K} + \frac{2pq s_{Dd} u_1}{Q}$	$\frac{2pq f_{Dd} a_2}{K} + \frac{2pq s_{Dd} u_2}{Q}$
D/D	$\frac{p^2 f_{DD} a_0}{K} + \frac{p^2 s_{DD} u_0}{Q}$	$\frac{p^2 f_{DD} a_1}{K} + \frac{p^2 s_{DD} u_1}{Q}$	$\frac{p^2 f_{DD} a_2}{K} + \frac{p^2 s_{DD} u_2}{Q}$

$h_{D2}h_{d1}$), $c_{d1} = 2n_0+n_1+n_3+(1-\alpha^4)n_4$, $c_{d2} = n_1+2n_2+n_5+\alpha^4 n_4$, $c_{D1} = n_3+2n_6+n_7+\alpha^4 n_4$, $c_{D2} = n_5+n_7+2n_8+(1-\alpha^4)n_4$. This yields overestimate of disease haplotypes. Conditioning on population disease allele frequency unchanged gives direct estimates $h_{d1} = h_{d1}q/sa$, $h_{d2} = h_{d2}q/sa$, $h_{D1} = h_{D1}p/su$, $h_{D2} = h_{D2}p/su$, where $sa = c_{d1} + c_{d2}$, $su = c_{D1} + c_{D2}$. The genotypic probabilities g are given in table A.7.

Table A.7: Genotypic probabilities for a disease locus and a marker

disease locus	marker 1/1	marker 1/2	marker 2/2
d/d	h_{d1}^2	$2h_{d1}h_{d2}$	h_{d2}^2
D/d	$2h_{D1}h_{d1}$	$2(h_{D1}h_{d2} + h_{D2}h_{d1})$	$2h_{D2}h_{d2}$
D/D	h_{D1}^2	$2\ h_{D1}h_{D2}$	h_{D2}^2

The updated table is now as in table A.8,

Table A.8: The updated count table for a disease locus and a marker

disease locus	marker 1/1	marker 1/2	marker 2/2
d/d	$\frac{g_0 f_{dd} a_0}{sa_1} + \frac{g_0 s_{dd} u_0}{su_1}$	$\frac{g_3 f_{Dd} a_0}{sa_1} + \frac{g_3 s_{Dd} u_0}{su_1}$	$\frac{g_6 f_{DD} a_0}{sa_1} + \frac{g_6 s_{DD} u_0}{su_1}$
D/d	$\frac{g_1 f_{dd} a_1}{sa_2} + \frac{g_1 s_{dd} u_1}{su_2}$	$\frac{g_4 f_{Dd} a_1}{sa_2} + \frac{g_4 s_{Dd} u_1}{su_2}$	$\frac{g_7 f_{DD} a_1}{sa_2} + \frac{g_7 s_{DD} u_1}{su_2}$
D/D	$\frac{g_2 f_{dd} a_2}{sa_3} + \frac{g_2 s_{dd} u_2}{su_3}$	$\frac{g_5 f_{Dd} a_2}{sa_3} + \frac{g_5 s_{Dd} u_2}{su_3}$	$\frac{g_8 f_{DD} a_2}{sa_3} + \frac{g_8 s_{DD} u_2}{su_3}$

where

$$
\begin{aligned}
sa_1 &= g_0 f_{dd} + g_3 f_{Dd} + g_6 f_{DD},\ su_1 = g_0 s_{dd} + g_3 s_{Dd} + g_6 s_{DD} \\
sa_2 &= g_1 f_{dd} + g_4 f_{Dd} + g_7 f_{DD},\ su_2 = g_1 s_{dd} + g_4 s_{Dd} + g_7 s_{DD} \\
sa_3 &= g_2 f_{dd} + g_5 f_{Dd} + g_8 f_{DD},\ su_3 = g_2 s_{dd} + g_5 s_{Dd} + g_8 s_{DD}
\end{aligned}
$$

and $sa = sa_1 + sa_2 + sa_3$, $su = su_1 + su_2 + su_3$ are the overall probabilities of being affected and unaffected.

The log-likelihood function is as follows.

$$\begin{aligned} l &= a_0 \ln((g_0 f_{dd} + g_3 f_{Dd} + g_6 f_{DD})/K_1) + u_0 \ln((g_0 s_{dd} + g_3 s_{Dd} + g_6 s_{DD})/K_2) \\ &+ a_1 \ln((g_1 f_{dd} + g_4 f_{Dd} + g_7 f_{DD})/K_1) + u_1 \ln((g_1 s_{dd} + g_4 s_{Dd} + g_7 s_{DD})/K_2) \\ &+ a_2 \ln((g_2 f_{dd} + g_5 f_{Dd} + g_8 f_{DD})/K_1) + u_2 \ln((g_2 s_{dd} + g_5 s_{Dd} + g_8 s_{DD})/K_2) \end{aligned}$$

where g's are the genotypic probabilities given above. K_1 and K_2 could either be sa and su or K and Q. We can iterate the counting procedure until successive changes in log-likelihoods is small than a predefined small constant. Different hypotheses can be tested using these haplotype frequency estimates. Denote the log-likelihood obtained from counting as l_1, that obtained from marker-marker association only as l_0, the log-likelihood ratio test statistic $2(l_1 - l_0)$ has asymptotic χ^2 distribution with 1 degree of freedom.

A disease locus and two biallelic markers

The initial table is as in table A.9.

Table A.9: Initial count table for a disease locus and two markers

disease locus	marker 1	marker 2 1/1	marker 2 1/2	marker 2 2/2
d/d	1/1	$\frac{q^2 f_{dd} a_0}{K} + \frac{q^2 s_{dd} u_0}{Q}$	$\frac{q^2 f_{dd} a_1}{K} + \frac{q^2 s_{dd} u_1}{Q}$	$\frac{q^2 f_{dd} a_2}{K} + \frac{q^2 s_{dd} u_2}{Q}$
	1/2	$\frac{q^2 f_{dd} a_3}{K} + \frac{q^2 s_{dd} u_3}{Q}$	$\frac{q^2 f_{dd} a_4}{K} + \frac{q^2 s_{dd} u_4}{Q}$	$\frac{q^2 f_{dd} a_5}{K} + \frac{q^2 s_{dd} u_5}{Q}$
	2/2	$\frac{q^2 f_{dd} a_6}{K} + \frac{q^2 s_{dd} u_6}{Q}$	$\frac{q^2 f_{dd} a_7}{K} + \frac{q^2 s_{dd} u_7}{Q}$	$\frac{q^2 f_{dd} a_8}{K} + \frac{q^2 s_{dd} u_8}{Q}$
D/d	1/1	$\frac{2pq f_{Dd} a_0}{K} + \frac{2pq s_{Dd} u_0}{Q}$	$\frac{2pq f_{Dd} a_1}{K} + \frac{2pq s_{Dd} u_1}{Q}$	$\frac{2pq f_{Dd} a_2}{K} + \frac{2pq s_{Dd} u_2}{Q}$
	1/2	$\frac{2pq f_{Dd} a_3}{K} + \frac{2pq s_{Dd} u_3}{Q}$	$\frac{2pq f_{Dd} a_4}{K} + \frac{2pq s_{Dd} u_4}{Q}$	$\frac{2pq f_{Dd} a_5}{K} + \frac{2pq s_{Dd} u_5}{Q}$
	2/2	$\frac{2pq f_{Dd} a_6}{K} + \frac{2pq s_{Dd} u_6}{Q}$	$\frac{2pq f_{Dd} a_7}{K} + \frac{2pq s_{Dd} u_7}{Q}$	$\frac{2pq f_{Dd} a_8}{K} + \frac{2pq s_{Dd} u_8}{Q}$
D/D	1/1	$\frac{p^2 f_{DD} a_0}{K} + \frac{p^2 s_{DD} u_0}{Q}$	$\frac{p^2 f_{DD} a_1}{K} + \frac{p^2 s_{DD} u_1}{Q}$	$\frac{p^2 f_{DD} a_2}{K} + \frac{p^2 s_{DD} u_2}{Q}$
	1/2	$\frac{p^2 f_{DD} a_3}{K} + \frac{p^2 s_{DD} u_3}{Q}$	$\frac{p^2 f_{DD} a_4}{K} + \frac{p^2 s_{DD} u_4}{Q}$	$\frac{p^2 f_{DD} a_5}{K} + \frac{p^2 s_{DD} u_5}{Q}$
	2/2	$\frac{p^2 f_{DD} a_6}{K} + \frac{p^2 s_{DD} u_6}{Q}$	$\frac{p^2 f_{DD} a_7}{K} + \frac{p^2 s_{DD} u_7}{Q}$	$\frac{p^2 f_{DD} a_8}{K} + \frac{p^2 s_{DD} u_8}{Q}$

The E-step is to obtain the expected probabilities

$$\begin{aligned} \alpha^4 &= h_{d11} h_{d22} / (h_{d11} h_{d22} + h_{d21} h_{d12}) \\ \alpha^{10} &= h_{D11} h_{d12} / (h_{D11} h_{d12} + h_{D12} h_{d11}) \\ \alpha^{12} &= h_{D11} h_{d21} / (h_{D11} h_{d21} + h_{D21} h_{d11}) \\ \alpha_1^{13} &= h_{D11} h_{d22} / (h_{D11} h_{d22} + h_{D12} h_{d21} + h_{D21} h_{d12} + h_{D22} h_{d11}) \end{aligned}$$

$$
\begin{aligned}
\alpha_2^{13} &= h_{D12}h_{d21}/(h_{D11}h_{d22}+h_{D12}h_{d21}+h_{D21}h_{d12}+h_{D22}h_{d11}) \\
\alpha_3^{13} &= h_{D21}h_{d12}/(h_{D11}h_{d22}+h_{D12}h_{d21}+h_{D21}h_{d12}+h_{D22}h_{d11}) \\
\alpha_4^{13} &= h_{D22}h_{d11}/(h_{D11}h_{d22}+h_{D12}h_{d21}+h_{D21}h_{d12}+h_{D22}h_{d11}) \\
\alpha^{14} &= h_{D12}h_{d22}/(h_{D12}h_{d22}+h_{D22}h_{d12}) \\
\alpha^{16} &= h_{D21}h_{d22}/(h_{D21}h_{d22}+h_{D22}h_{d21}) \\
\alpha^{22} &= h_{D11}h_{D22}/(h_{D11}h_{D22}+h_{D21}h_{D12})
\end{aligned}
$$

and the haplotype counts

$$
\begin{aligned}
c_{d11} &= 2n_0+n_1+n_3+\alpha^4 n_4+n_9+(1-\alpha^{10})n_{10}+(1-\alpha^{12})n_{12}+\alpha_4^{13}n_{13} \\
c_{d12} &= n_1+2n_2+n_5+(1-\alpha^4)n_4+\alpha^{10}n_{10}+n_{11}+\alpha_3^{13}n_{13}+(1-\alpha^{14})n_{14} \\
c_{d21} &= n_3+2n_6+n_7+(1-\alpha^4)n_4+\alpha^{12}n_{12}+\alpha_2^{13}n_{13}+n_{15}+(1-\alpha^{16})n_{16} \\
c_{d22} &= n_5+n_7+2n_8+\alpha^4 n_4+\alpha_1^{13}n_{13}+\alpha^{14}n_{14}+\alpha^{16}n_{16}+n_{17} \\
c_{D11} &= n_9+\alpha^{10}n_{10}+\alpha^{12}n_{12}+\alpha_1^{13}n_{13}+2n_{18}+n_{19}+n_{21}+\alpha^{22}n_{22} \\
c_{D12} &= (1-\alpha^{10})n_{10}+n_{11}+\alpha_2^{13}n_{13}+\alpha^{14}n_{14}+n_{19}+2n_{20}+(1-\alpha^{22})n_{22}+n_{23} \\
c_{D21} &= (1-\alpha^{12})n_{12}+\alpha_3^{13}n_{13}+n_{15}+\alpha^{16}n_{16}+n_{21}+(1-\alpha^{22})n_{22}+2n_{24}+n_{25} \\
c_{D22} &= \alpha_4^{13}n_{13}+(1-\alpha^{14})n_{14}+(1-\alpha^{16})n_{16}+n_{17}+\alpha^{22}n_{22}+n_{23}+n_{25}+2n_{26}
\end{aligned}
$$

By conditioning, $h_{d11}=qh_{d11}/sa,\ h_{D11}=ph_{D11}/su,\ h_{d12}=qh_{d12}/sa,\ h_{D12}=ph_{D12}/su,$ $h_{d21}=qh_{d21}/sa,\ h_{D21}=ph_{D21}/su,\ h_{d22}=qh_{d22}/sa,\ h_{D22}=ph_{D22}/su,$ where $sa=c_{d11}+c_{d12}+c_{d21}+c_{d22}$, $su=c_{D11}+c_{D12}+c_{D21}+c_{D22}$. We have the genotypic probabilities, g_i, $i=0,\ldots,26$ in Table A.10.

Table A.10: The genotypic probabilities for a disease locus and two markers

disease		marker 2		
locus	marker 1	1/1	1/2	2/2
d/d	1/1	h_{d11}^2	$2h_{d12}h_{d11}$	h_{d12}^2
	1/2	$2h_{d11}h_{d21}$	$2(h_{d12}h_{d21}+h_{d11}h_{d22})$	$2h_{d12}h_{d22}$
	2/2	h_{d21}^2	$2h_{d21}h_{d22}$	h_{d22}^2
D/d	1/1	$2h_{D11}h_{d11}$	$2(h_{D11}h_{d12}+h_{D12}h_{d11})$	$2h_{D12}h_{d12}$
	1/2	$2h_{D11}h_{d21}$ $+2h_{D21}h_{d11}$	$2(h_{D11}h_{d22}+h_{D12}h_{d21})$ $+2(h_{D21}h_{d12}+h_{D22}h_{d11})$	$2h_{D12}h_{d22}$ $+2h_{D22}h_{d12}$
	2/2	$2h_{D21}h_{d21}$	$2(h_{D21}h_{d22}+h_{D22}h_{d21})$	$2h_{D22}h_{d22}$
D/D	1/1	h_{D11}^2	$2h_{D11}h_{D12}$	h_{D12}^2
	1/2	$2h_{D11}h_{D21}$	$2(h_{D11}h_{D22}+h_{D21}h_{D12})$	$2h_{D12}h_{D22}$
	2/2	h_{D21}^2	$2h_{D21}h_{D22}$	h_{D22}^2

Now our gene counting table turns out to be Table A.11, where

$$
\begin{aligned}
sa_1 &= g_0 f_{dd} + g_9 f_{Dd} + g_{18} f_{DD},\ su_1 = g_0 s_{dd} + g_9 s_{Dd} + g_{18} s_{DD} \\
sa_2 &= g_1 f_{dd} + g_{10} f_{Dd} + g_{19} f_{DD},\ su_2 = g_1 s_{dd} + g_{10} s_{Dd} + g_{19} s_{DD} \\
sa_3 &= g_2 f_{dd} + g_{11} f_{Dd} + g_{20} f_{DD},\ su_3 = g_2 s_{dd} + g_{11} s_{Dd} + g_{20} s_{DD} \\
sa_4 &= g_3 f_{dd} + g_{12} f_{Dd} + g_{21} f_{DD},\ su_4 = g_3 s_{dd} + g_{12} s_{Dd} + g_{21} s_{DD} \\
sa_5 &= g_4 f_{dd} + g_{13} f_{Dd} + g_{22} f_{DD},\ su_5 = g_4 s_{dd} + g_{13} s_{Dd} + g_{22} s_{DD} \\
sa_6 &= g_5 f_{dd} + g_{14} f_{Dd} + g_{23} f_{DD},\ su_6 = g_5 s_{dd} + g_{14} s_{Dd} + g_{23} s_{DD} \\
sa_7 &= g_5 f_{dd} + g_{15} f_{Dd} + g_{24} f_{DD},\ su_7 = g_5 s_{dd} + g_{15} s_{Dd} + g_{24} s_{DD} \\
sa_8 &= g_7 f_{dd} + g_{16} f_{Dd} + g_{25} f_{DD},\ su_8 = g_7 s_{dd} + g_{16} s_{Dd} + g_{25} s_{DD} \\
sa_9 &= g_8 f_{dd} + g_{17} f_{Dd} + g_{26} f_{DD},\ su_9 = g_8 s_{dd} + g_{17} s_{Dd} + g_{26} s_{DD}
\end{aligned}
$$

Table A.11: The updated gene count table for a disease locus and two markers

disease locus	marker 1	marker 2		
		1/1	1/2	2/2
d/d	1/1	$\frac{g_0 f_{dd} a_0}{sa_1} + \frac{g_0 s_{dd} u_0}{su_1}$	$\frac{g_1 f_{dd} a_1}{sa_2} + \frac{g_1 s_{dd} u_1}{su_2}$	$\frac{g_2 f_{dd} a_2}{sa_3} + \frac{g_2 s_{dd} u_2}{su_3}$
	1/2	$\frac{g_3 f_{dd} a_3}{sa_4} + \frac{g_3 s_{dd} u_3}{su_4}$	$\frac{g_4 f_{dd} a_4}{sa_5} + \frac{g_4 s_{dd} u_4}{su_5}$	$\frac{g_5 f_{dd} a_5}{sa_6} + \frac{g_5 s_{dd} u_5}{su_6}$
	2/2	$\frac{g_6 f_{dd} a_6}{sa_7} + \frac{g_6 s_{dd} u_6}{su_7}$	$\frac{g_7 f_{dd} a_7}{sa_8} + \frac{g_7 s_{dd} u_7}{su_8}$	$\frac{g_8 f_{dd} a_8}{sa_9} + \frac{g_8 s_{dd} u_8}{su_9}$
D/d	1/1	$\frac{g_9 f_{Dd} a_0}{sa_1} + \frac{g_9 s_{Dd} u_0}{su_1}$	$\frac{g_{10} f_{Dd} a_1}{sa_2} + \frac{g_{10} s_{Dd} u_1}{su_2}$	$\frac{g_{11} f_{Dd} a_2}{sa_3} + \frac{g_{11} s_{Dd} u_2}{su_3}$
	1/2	$\frac{g_{12} f_{Dd} a_3}{sa_4} + \frac{g_{12} s_{Dd} u_3}{su_4}$	$\frac{g_{13} f_{Dd} a_4}{sa_5} + \frac{g_{13} s_{Dd} u_4}{su_5}$	$\frac{g_{14} f_{Dd} a_5}{sa_6} + \frac{g_{14} s_{Dd} u_5}{su_6}$
	2/2	$\frac{g_{15} f_{Dd} a_6}{sa_7} + \frac{g_{15} s_{Dd} u_6}{su_7}$	$\frac{g_{16} f_{Dd} a_7}{sa_8} + \frac{g_{16} s_{Dd} u_7}{su_8}$	$\frac{g_{17} f_{Dd} a_8}{sa_9} + \frac{g_{17} s_{Dd} u_8}{su_9}$
D/D	1/1	$\frac{g_{18} f_{DD} a_0}{sa_1} + \frac{g_{18} s_{DD} u_0}{su_1}$	$\frac{g_{19} f_{DD} a_1}{sa_2} + \frac{g_{19} s_{DD} u_1}{su_2}$	$\frac{g_{20} f_{DD} a_2}{sa_3} + \frac{g_{20} s_{DD} u_2}{su_3}$
	1/2	$\frac{g_{21} f_{DD} a_3}{sa_4} + \frac{g_{21} s_{DD} u_3}{su_4}$	$\frac{g_{22} f_{DD} a_4}{sa_5} + \frac{g_{22} s_{DD} u_4}{su_5}$	$\frac{g_{23} f_{DD} a_5}{sa_6} + \frac{g_{23} s_{DD} u_5}{su_6}$
	2/2	$\frac{g_{24} f_{DD} a_6}{sa_7} + \frac{g_{24} s_{DD} u_6}{su_7}$	$\frac{g_{25} f_{DD} a_7}{sa_8} + \frac{g_{25} s_{DD} u_7}{su_8}$	$\frac{g_{26} f_{DD} a_8}{sa_9} + \frac{g_{26} s_{DD} u_8}{su_9}$

We can calculate the probabilities of being affected and unaffected based on our sample $sa = \sum_{i=1}^{9} sa_i,\ su = \sum_{i=1}^{9} su_i$. The log-likelihood function can be obtained in a similar fashion.

$$
\begin{aligned}
l &= a_0 \ln((g_0 f_{dd} + g_9 f_{Dd} + g_{18} f_{DD})/K_1) + u_0 \ln((g_0 s_{dd} + g_9 s_{Dd} + g_{18} s_{DD})/K_2) \\
&+ a_1 \ln((g_1 f_{dd} + g_{10} f_{Dd} + g_{19} f_{DD})/K_1) + u_1 \ln((g_1 s_{dd} + g_{10} s_{Dd} + g_{19} s_{DD})/K_2) \\
&+ a_2 \ln((g_2 f_{dd} + g_{11} f_{Dd} + g_{20} f_{DD})/K_1) + u_2 \ln((g_2 s_{dd} + g_{11} s_{Dd} + g_{20} s_{DD})/K_2) \\
&+ a_3 \ln((g_3 f_{dd} + g_{12} f_{Dd} + g_{21} f_{DD})/K_1) + u_3 \ln((g_3 s_{dd} + g_{12} s_{Dd} + g_{21} s_{DD})/K_2) \\
&+ a_4 \ln((g_4 f_{dd} + g_{13} f_{Dd} + g_{22} f_{DD})/K_1) + u_4 \ln((g_4 s_{dd} + g_{13} s_{Dd} + g_{22} s_{DD})/K_2) \\
&+ a_5 \ln((g_5 f_{dd} + g_{14} f_{Dd} + g_{23} f_{DD})/K_1) + u_5 \ln((g_5 s_{dd} + g_{14} s_{Dd} + g_{23} s_{DD})/K_2)
\end{aligned}
$$

$$+ \quad a_6 \ln((g_5 f_{dd} + g_{15} f_{Dd} + g_{24} f_{DD})/K_1) + u_6 \ln((g_5 s_{dd} + g_{15} s_{Dd} + g_{24} s_{DD})/K_2)$$
$$+ \quad a_7 \ln((g_7 f_{dd} + g_{16} f_{Dd} + g_{25} f_{DD})/K_1) + u_7 \ln((g_7 s_{dd} + g_{16} s_{Dd} + g_{25} s_{DD})/K_2)$$
$$+ \quad a_8 \ln((g_8 f_{dd} + g_{17} f_{Dd} + g_{26} f_{DD})/K_1) + u_8 \ln((g_8 s_{dd} + g_{17} s_{Dd} + g_{26} s_{DD})/K_2)$$

where K_1 and K_2 could either be sa and su or K and Q. The procedure above can be iterated to pre-specified criteria based on the log-likelihoods.

A disease locus with three alleles

Denote alleles of the disease locus to be d_1, d_2, d_3, their frequencies p, q, r, penetrances f_{ij}, s_{ij}, $i, j = 1, \ldots, 3$, our observed data can be decomposed into table A.12.

Table A.12: Gene counts for a triallelic disease locus and two biallelic markers

disease		marker 2		
locus	marker 1	1/1	1/2	2/2
	1/1	n_0	n_1	n_2
d_1/d_1	1/2	n_3	n_4	n_5
	2/2	n_6	n_7	n_8
	1/1	n_9	n_{10}	n_{11}
d_1/d_2	1/2	n_{12}	n_{13}	n_{14}
	2/2	n_{15}	n_{16}	n_{17}
	1/1	n_{18}	n_{19}	n_{20}
d_2/d_2	1/2	n_{21}	n_{22}	n_{23}
	2/2	n_{24}	n_{25}	n_{26}
	1/1	n_{27}	n_{28}	n_{29}
d_1/d_3	1/2	n_{30}	n_{31}	n_{32}
	2/2	n_{33}	n_{34}	n_{35}
	1/1	n_{36}	n_{37}	n_{38}
d_2/d_3	1/2	n_{39}	n_{40}	n_{41}
	2/2	n_{42}	n_{43}	n_{44}
	1/1	n_{45}	n_{46}	n_{47}
d_3/d_3	1/2	n_{48}	n_{49}	n_{50}
	2/2	n_{51}	n_{52}	n_{53}

Cells in the initial table can be one of the six forms.

$$\frac{p^2 f_{11} a_i}{K} + \frac{p^2 s_{11} u_i}{Q}, \qquad i = 0, \ldots, 9$$
$$\frac{2pq f_{12} a_i}{K} + \frac{2pq s_{12} u_i}{Q}, \qquad i = 10, \ldots, 17$$
$$\frac{q^2 f_{22} a_i}{K} + \frac{q^2 s_{22} u_i}{Q}, \qquad i = 18, \ldots, 26$$

$$
\begin{aligned}
\frac{2prf_{13}a_i}{K} + \frac{2prs_{13}u_i}{Q}, &\qquad i = 27, \ldots, 35 \\
\frac{2qrf_{23}a_i}{K} + \frac{2qrs_{23}u_i}{Q}, &\qquad i = 36, \ldots, 47 \\
\frac{r^2 f_{33}a_i}{K} + \frac{r^2 s_{33}u_i}{Q}, &\qquad i = 45, \ldots, 53
\end{aligned}
$$

The haplotype counts are obtained as follows

$$
\begin{aligned}
c_{d_1 11} &= 2n_0 + n_1 + n_3 + \alpha^4 n_4 \\
&+ n_9 + (1-\alpha^{10})n_{10} + (1-\alpha^{12})n_{12} + \alpha_4^{13} n_{13} \\
&+ n_{27} + (1-\alpha^{28})n_{28} + (1-\alpha^{30})n_{30} + \alpha_4^{31} n_{31} \\
c_{d_1 12} &= n_1 + 2n_2 + (1-\alpha^4)n_4 + n_5 \\
&+ \alpha^{10} n_{10} + n_{11} + \alpha_3^{13} n_{13} + (1-\alpha^{14})n_{14} \\
&+ \alpha^{28} n_{28} + n_{29} + \alpha_3^{31} n_{31} + (1-\alpha^{32})n_{32} \\
c_{d_1 21} &= n_3 + (1-\alpha^4)n_4 + 2n_6 + n_7 \\
&+ \alpha^{12} n_{12} + \alpha_2^{13} n_{13} + n_{15} + (1-\alpha^{16})n_{16} \\
&+ \alpha^{30} n_{30} + \alpha_2^{31} n_{31} + n_{33} + (1-\alpha^{34})n_{34} \\
c_{d_1 22} &= \alpha^4 n_4 + n_5 + n_7 + 2n_8 \\
&+ \alpha_1^{13} n_{13} + \alpha^{14} n_{14} + \alpha^{16} n_{16} + n_{17} \\
&+ \alpha_1^{31} n_{31} + \alpha^{32} n_{32} + \alpha^{34} n_{34} + n_{35} \\
c_{d_2 11} &= 2n_{18} + n_{19} + n_{24} + \alpha^{25} n_{25} \\
&+ n_9 + \alpha^{10} n_{10} + \alpha^{12} n_{12} + \alpha_1^{13} n_{13} \\
&+ n_{36} + (1-\alpha^{37})n_{37} + (1-\alpha^{39})n_{39} + \alpha_4^{40} n_{40} \\
c_{d_2 12} &= n_{22} + 2n_{23} + (1-\alpha^{25})n_{25} + n_{26} \\
&+ (1-\alpha^{10})n_{10} + n_{11} + \alpha_2^{13} n_{13} + \alpha^{14} n_{14} \\
&+ \alpha^{37} n_{37} + n_{38} + \alpha_3^{40} n_{40} + (1-\alpha^{41})n_{41} \\
c_{d_2 21} &= n_{24} + (1-\alpha^{25})n_{25} + 2n_{27} + n_{28} \\
&+ (1-\alpha^{12})n_{12} + \alpha_3^{13} n_{13} + n_{15} + \alpha^{16} n_{16} \\
&+ \alpha^{39} n_{39} + \alpha_2^{40} n_{40} + n_{42} + (1-\alpha^{43})n_{43} \\
c_{d_2 22} &= \alpha^{25} n_{25} + n_{26} + n_{28} + 2n_{29} \\
&+ \alpha_4^{13} n_{13} + (1-\alpha^{14})n_{14} + (1-\alpha^{16})n_{16} + n_{17} \\
&+ \alpha_1^{40} n_{40} + \alpha^{41} n_{41} + \alpha^{43} n_{43} + n_{44}
\end{aligned}
$$

$$
\begin{aligned}
c_{d_3 11} &= 2n_{45} + n_{46} + n_{47} + \alpha^{49} n_{49} \\
&+ n_{27} + \alpha^{28} n_{28} + \alpha^{30} n_{30} + \alpha_1^{31} n_{31} \\
&+ n_{36} + \alpha^{37} n_{37} + \alpha^{39} n_{39} + \alpha_1^{40} n_{40} \\
c_{d_3 12} &= n_{46} + 2n_{47} + (1 - \alpha^{49}) n_{49} + n_{50} \\
&+ (1 - \alpha^{28}) n_{28} + n_{29} + \alpha_2^{31} n_{31} + \alpha^{32} n_{32} \\
&+ (1 - \alpha^{37}) n_{37} + n_{38} + \alpha_2^{40} n_{40} + \alpha^{41} n_{41} \\
c_{d_3 21} &= n_{48} + (1 - \alpha^{49}) n_{49} + 2n_{51} + n_{52} \\
&+ (1 - \alpha^{30}) n_{30} + \alpha_3^{31} n_{31} + n_{33} + \alpha^{34} n_{34} \\
&+ (1 - \alpha^{39}) n_{39} + \alpha_3^{40} n_{40} + n_{42} + \alpha^{43} n_{43} \\
c_{d_3 22} &= \alpha^{49} n_{49} + n_{50} + n_{52} + 2n_{53} \\
&+ \alpha_4^{31} n_{31} + (1 - \alpha^{32}) n_{32} + (1 - \alpha^{34}) n_{34} + n_{35} \\
&+ \alpha_4^{40} n_{40} + (1 - \alpha^{41}) n_{41} + (1 - \alpha^{43}) n_{43} + n_{44}
\end{aligned}
$$

for example $\alpha^4 = h_{d_1 11} h_{d_1 22} / (h_{d_1 11} h_{d_1 22} + h_{d_1 12} h_{d_1 21})$, The log-likelihood function is easily written as follows.

$$
\begin{aligned}
l &= a_0 \ln((g_0 f_{11} + g_9 f_{12} + g_{18} f_{22} + g_{27} f_{13} + g_{36} f_{23} + g_{45} f_{33})/K_1) \\
&+ u_0 \ln((g_0 s_{11} + g_9 s_{12} + g_{18} s_{22} + g_{27} s_{13} + g_{36} s_{23} + g_{45} s_{33})/K_2) \\
&+ a_1 \ln((g_1 f_{11} + g_{10} f_{12} + g_{19} f_{22} + g_{28} f_{13} + g_{37} f_{23} + g_{46} f_{33})/K_1) \\
&+ u_1 \ln((g_1 s_{11} + g_{10} s_{12} + g_{19} s_{22} + g_{28} s_{13} + g_{37} s_{23} + g_{46} s_{33})/K_2) \\
&+ a_2 \ln((g_2 f_{11} + g_{11} f_{12} + g_{20} f_{22} + g_{29} f_{13} + g_{38} f_{23} + g_{47} f_{33})/K_1) \\
&+ u_2 \ln((g_2 s_{11} + g_{11} s_{12} + g_{20} s_{22} + g_{29} s_{13} + g_{38} s_{23} + g_{47} s_{33})/K_2) \\
&+ a_3 \ln((g_3 f_{11} + g_{12} f_{12} + g_{21} f_{22} + g_{30} f_{13} + g_{39} f_{23} + g_{48} f_{33})/K_1) \\
&+ u_3 \ln((g_3 s_{11} + g_{12} s_{12} + g_{21} s_{22} + g_{30} s_{13} + g_{39} s_{23} + g_{48} s_{33})/K_2) \\
&+ a_4 \ln((g_4 f_{11} + g_{13} f_{12} + g_{22} f_{22} + g_{31} f_{13} + g_{40} f_{23} + g_{49} f_{33})/K_1) \\
&+ u_4 \ln((g_4 s_{11} + g_{13} s_{12} + g_{22} s_{22} + g_{31} s_{13} + g_{40} s_{23} + g_{49} s_{33})/K_2) \\
&+ a_5 \ln((g_5 f_{11} + g_{14} f_{12} + g_{23} f_{22} + g_{32} f_{13} + g_{41} f_{23} + g_{50} f_{33})/K_1) \\
&+ u_5 \ln((g_5 s_{11} + g_{14} s_{12} + g_{23} s_{22} + g_{32} s_{13} + g_{41} s_{23} + g_{50} s_{33})/K_2) \\
&+ a_6 \ln((g_5 f_{11} + g_{15} f_{12} + g_{24} f_{22} + g_{33} f_{13} + g_{42} f_{23} + g_{51} f_{33})/K_1) \\
&+ u_6 \ln((g_5 s_{11} + g_{15} s_{12} + g_{24} s_{22} + g_{33} s_{13} + g_{42} s_{23} + g_{51} s_{33})/K_2) \\
&+ a_7 \ln((g_7 f_{11} + g_{16} f_{12} + g_{25} f_{22} + g_{34} f_{13} + g_{43} f_{23} + g_{52} f_{33})/K_1)
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad u_7 \ln((g_7 s_{11} + g_{16} s_{12} + g_{25} s_{22} + g_{34} s_{13} + g_{43} s_{23} + g_{52} s_{33})/K_2) \\
&+ \quad a_8 \ln((g_8 f_{11} + g_{17} f_{12} + g_{26} f_{22} + g_{35} f_{13} + g_{44} f_{23} + g_{53} f_{33})/K_1) \\
&+ \quad u_8 \ln((g_8 s_{11} + g_{17} s_{12} + g_{26} s_{22} + g_{35} s_{13} + g_{44} s_{23} + g_{53} s_{33})/K_2)
\end{aligned}
$$

where K_1 and K_2 could either be sa and su or K and Q. $sa = \sum_{i=1}^{9} sa_i$, $su = \sum_{i=1}^{9} su_i$ with

$$
\begin{aligned}
sa_1 &= g_0 f_{11} + g_9 f_{12} + g_{18} f_{22} + g_{18} f_{22} + g_{27} f_{13} + g_{36} f_{23} + g_{45} f_{33} \\
sa_2 &= g_1 f_{11} + g_{10} f_{12} + g_{19} f_{22} + g_{18} s_{22} + g_{27} s_{13} + g_{36} s_{23} + g_{45} s_{33} \\
sa_3 &= g_2 f_{11} + g_{11} f_{12} + g_{20} f_{22} + g_{19} f_{22} + g_{28} f_{13} + g_{37} f_{23} + g_{46} f_{33} \\
sa_4 &= g_3 f_{11} + g_{12} f_{12} + g_{21} f_{22} + g_{19} s_{22} + g_{28} s_{13} + g_{37} s_{23} + g_{46} s_{33} \\
sa_5 &= g_4 f_{11} + g_{13} f_{12} + g_{22} f_{22} + g_{20} f_{22} + g_{29} f_{13} + g_{38} f_{23} + g_{47} f_{33} \\
sa_6 &= g_5 f_{11} + g_{14} f_{12} + g_{23} f_{22} + g_{20} s_{22} + g_{29} s_{13} + g_{38} s_{23} + g_{47} s_{33} \\
sa_7 &= g_5 f_{11} + g_{15} f_{12} + g_{24} f_{22} + g_{21} f_{22} + g_{30} f_{13} + g_{39} f_{23} + g_{48} f_{33} \\
sa_8 &= g_7 f_{11} + g_{16} f_{12} + g_{25} f_{22} + g_{21} s_{22} + g_{30} s_{13} + g_{39} s_{23} + g_{48} s_{33} \\
sa_9 &= g_8 f_{11} + g_{17} f_{12} + g_{26} f_{22} + g_{22} f_{22} + g_{31} f_{13} + g_{40} f_{23} + g_{49} f_{33} \\
su_1 &= g_0 s_{11} + g_9 s_{12} + g_{18} s_{22} + g_{22} s_{22} + g_{31} s_{13} + g_{40} s_{23} + g_{49} s_{33} \\
su_2 &= g_1 s_{11} + g_{10} s_{12} + g_{19} s_{22} + g_{23} f_{22} + g_{32} f_{13} + g_{41} f_{23} + g_{50} f_{33} \\
su_3 &= g_2 s_{11} + g_{11} s_{12} + g_{20} s_{22} + g_{23} s_{22} + g_{32} s_{13} + g_{41} s_{23} + g_{50} s_{33} \\
su_4 &= g_3 s_{11} + g_{12} s_{12} + g_{21} s_{22} + g_{24} f_{22} + g_{33} f_{13} + g_{42} f_{23} + g_{51} f_{33} \\
su_5 &= g_4 s_{11} + g_{13} s_{12} + g_{22} s_{22} + g_{24} s_{22} + g_{33} s_{13} + g_{42} s_{23} + g_{51} s_{33} \\
su_6 &= g_5 s_{11} + g_{14} s_{12} + g_{23} s_{22} + g_{25} f_{22} + g_{34} f_{13} + g_{43} f_{23} + g_{52} f_{33} \\
su_7 &= g_5 s_{11} + g_{15} s_{12} + g_{24} s_{22} + g_{25} s_{22} + g_{34} s_{13} + g_{43} s_{23} + g_{52} s_{33} \\
su_8 &= g_7 s_{11} + g_{16} s_{12} + g_{25} s_{22} + g_{26} f_{22} + g_{35} f_{13} + g_{44} f_{23} + g_{53} f_{33} \\
su_9 &= g_8 s_{11} + g_{17} s_{12} + g_{26} s_{22} + g_{26} s_{22} + g_{35} s_{13} + g_{44} s_{23} + g_{53} s_{33}
\end{aligned}
$$

It is clear that the gene counting algorithm will soon become unwieldy and a computer algorithm is more appropriate.

General case and further remarks

EH and **EHPLUS** implement algorithms for three sampling schemes: a single phenotypic

sample, a population sample of cases and controls, and a randomly ascertained sample of cases and control. Following Ott, in his 1998 version of **EH** documentation, schemes involving m markers, each with alleles a_j, $j = 1, \ldots, m$, can be briefly described as follows.

Case 1. A group of subjects with genotypic information M_i, $i = 1, \ldots, N$ with log-likelihood function $\sum_{i=1}^{N} \ln P(M_i) = \sum_{i=1}^{n} \ln \sum_g P(M_i|g)P(g)$, where N is the sample size, g is a multilocus genotype at all loci, $P(M_i|g)$ is the product of penetrances taking values of 0 or 1, and $P(g)$ the genotype probability based on the population haplotype frequencies assuming random union of haplotypes. The number of these genotypes is $k(k+1)/2$, with k being the number of haplotypes. The original data is conveniently organised into an m-dimensional contingency table, with each dimension being indexed by unique genotype identifiers $l + u(u-1)/2$, l and u are alleles at marker i, $l \leq u$, $i = 1, \ldots, m$. The log-likelihood can simply be obtained from the observed multilocus genotype counts and multinomial probabilities. Each population haplotype frequency is simply the product of all constituent allele frequencies assuming no association between loci.

Case 2. A random population sample of n_A cases and n_U controls with unconditional log-likelihood of all the phenotypic data $\sum_{i=1}^{n_A} \ln P(A, M_i) + \sum_{j=1}^{n_U} \ln P(U, M_j)$, where $P(A, M_i) = P(A, M_i|g)P(g)$ and $P(U, M_j) = P(U, M_j|g)P(g)$, while $P(A, M_i|g)$ and $P(U, M_j|g)$ are given by the product of all penetrances at individual loci. $P(g)$ is the genotype probability incorporating disease locus.

Case 3. Two independently ascertained samples of cases and controls with associated conditional likelihoods $P(M|A) = \sum_g P(M|g, A)\ P(A|g)P(g)/P(A)$, $P(M|U) = \sum_g P(M|g, U)\ P(U|g)P(g)/P(U)$, respectively. We have $P(A) = \sum_g P(A|g)P(g)$, $P(U) = \sum_g P(U|g)P(g)$, i.e., summing over all the possible genotype configurations, $P(g)$ being given by the population haplotype frequencies. Each cell in the contingency table will contribute probabilities $P(A|g)P(g)$ of being affected and $P(U|g)P(g)$ being unaffected to $P(A)$ and $P(U)$, where $P(A|g) + P(U|g) = 1$. However the cases are over-represented in the sample the contributions are scaled by $P(A)$ and $P(U)$ to have total cell probabilities sum to 1, $P(A|g)P(g)/P(A) = P(g|A)$, $P(U|g)P(g)/P(U) = P(g|U)$. The conditional log-likelihood can be expressed from its unconditional counterpart with an extra term of $n_A \ln P(A) + n_U \ln P(U)$.

The MLEs of haplotype frequencies is furnished by gene counting. Starting from a set of initial haplotype frequencies, a new set of frequencies is obtained, $P(h|A)$ for cases and

$P(h|U)$ for control, such that $P(h) = P(h|A)P(A) + P(h|U)P(U)$, which serves as initial values for the next iteration. Specifically, the gene counts are estimated from an iteratively updated table with $L = 3\prod_{j=1}^{m} a_j(a_j+1)/2$ cells obtained by conditioning marker genotypes on disease locus. Repeating the procedure would suffice to get the MLEs.

Three hypotheses H_0, H_1, H_2, of allelic association can be considered. H_0 assumes no marker-marker or disease-marker association so haplotype frequencies are simply product of individual disease and marker allele frequencies. H_1 only allows for marker-marker association so haplotype frequencies are estimated from pooled case-control sample with disease locus as an extra locus. H_2 imposes a disease model to the whole sample and get haplotype frequency estimates by maximising the log-likelihood function. Appropriate log-likelihood ratio statistics can be constructed.

Multiple samples can also be subject to heterogeneity analysis (Workman & Niswander 1970). Due to the large number of degrees of freedom, empirical p value can then be obtained by permutation. Confidence interval or threshold can be based on binomial distribution when $Np \geq 5$ (N is the number of replicates) and there is little need for a confidence interval when p value is small (Nettleton & Doerge 2000). The standard error of p can also be estimated by subdividing the number of replicates into B batches (Hastings 1970; Guo & Thompson 1992), a p value is calculated separately for each batch as $p_i, i = 1, \ldots, B$, the standard error of p is then $\sqrt{\sum_{i=1}^{B}(p-p_i)/B(B-1)}$. The process is stopped when the estimated standard error is smaller than a predefined value specified by the user.

The method of maximum likelihood (Thompson et al. 1988; Cox et al. 1998) can provide asymptotic variances but the MLEs are more difficult to obtain. Recall that the allele frequencies for the two markers are p, $1-p$ and q, $1-q$, with observed counts n_i for joint genotypes g_i, $i = 0, \ldots, 8$, with $\sum_{i=0}^{8} n_i = N$. Rewritten in terms of disequilibrium parameter D, haplotype frequencies are now $h_{11} = pq + D$, $h_{12} = p(1-q) - D$, $h_{21} = (1-p)q - D$, $h_{22} = (1-p)(1-q) + D$. Let $\theta = (\theta_1, \theta_2, \theta_3) = (p, q, D)^T$ be the vector of parameters, the likelihood and log-likelihood are given by $\left(n!/\prod_{i=0}^{8} n_i!\right)\prod_{i=0}^{8} g_i^{n_i}$ and constant $+ \sum_{i=0}^{8} n_i \ln(g_i)$, respectively. The variance-covariance matrix of θ can be approximated by $V(\theta) = I^{-1}(\theta)$, where $I(\theta) = \left[E\left(-\frac{\partial^2 l(\theta)}{\partial\theta_j \partial\theta_k}\right)\right]$ is the information matrix with (j,k) element being $N\left[\sum_{i=0}^{8} \frac{\partial y_i}{\partial\theta_j}\frac{\partial y_i}{\partial\theta_k} y_i^{-1}\right]$, $j, k = 0, \ldots, 2$. The variance of D for N observations evaluated at $D = 0$ is $V_N(D=0) = p(1-p)q(1-q)/N$ and from which the power can be calculated. The critical region for rejection of null hypothesis at 5% significance level is

$|\hat{D}| > 1.96\sqrt{V_N(D=0)}$, and the power is $w = P(\text{reject } H_0 : D = 0 | D \neq 0)$, i.e.,

$$w = \Phi\left(\frac{-1.96\sqrt{V_N(D=0)} + D}{\sqrt{V_N(D)}}\right) + \Phi\left(\frac{-1.96\sqrt{V_N(D=0)} - D}{\sqrt{V_N(D)}}\right)$$

where $\Phi(.)$ is the standard normal distribution function. When $D \neq 0$, $V_N(D)$ is calculated numerically. An approximation to the required sample size is obtained by taking w to be the first rhs term for negative D and the second rhs term for positive D. In either case, $N = \left(\Phi^{-1}(w) + k(D)/m(D)\right)^2$, where $k(D) = 1.96\sqrt{p(1-p)q(1-q)}/\sqrt{V_1(D)}$, $m(D) = D/\sqrt{V_1(D)}$ with $V_1(D) = I_{2,2}^{-1}(\theta)$ being the single-observation variance based on the (2,2) element of the information matrix.

The quantity $\rho^2 = D^2/[p(1-p)q(1-q)]$ is of special interests, in that $2N\rho^2$ can either be compared to χ^2 with 1 degree of freedom, or ρ be treated as correlation coefficient for normal test using Fisher's z-transformation (Weir 1996). There is a simple relationship $E(\rho^2) \approx 1/(1+4N_e\theta)$, where N_e is the effective population size and θ is the recombination rate (Hill & Robertson 1968; Sved 1971). Hill (1974) extended the approach in Hill & Robertson (1968) to three or more loci.

A scaled LD measure is $D' = D/D_{max}$, where

$$D_{max} = \begin{cases} \min(pq, (1-p)(1-q)), & D < 0 \\ \min(p(1-q), q(1-p)), & D > 0 \end{cases}$$

Zapata et al. (1997) gave its standard error estimates. A global disequilibrium statistic based on pairwise disequilibrium measure $D_{ij} = h_{ij} - p_i q_j$ can be defined as $W = \sqrt{\sum_i \sum_j D_{ij}^2 (p_i q_j)^{-1}}$, where p_i and q_j, $i = 1, \ldots, k$; $j = 1, \ldots, l$ are the observed allele frequencies at each of two loci having k and l alleles. A normalised version $W' = W/\sqrt{\min(k,l) - 1}$ lies in [0,1]. If the sample size is N, then NW^2 can be used for testing significance of the global disequilibrium and refer to χ^2 with degrees of freedom $(k-1)(l-1)$ (Klitz et al. 1995; Long et al. 1995). There is also counterpart of the two biallelic D' to two multialleic markers, and Zapata et al. (2001) gave its standard error.

A permutation-based linkage disequilibrium estimate (Zhao H et al. 1999) is defined as $\hat{\xi} = \sqrt{2f}\,(t - \mu/\sigma)\,/N$, where t is log-likelihood ratio test statistic from the observed data, f its degrees of freedom, and N the number of replicates. The mean (μ) and variance (σ^2) of the likelihood ratio test statistic are based on its empirical distribution obtained

by simulation. The sample variance of $\hat{\xi}$ is $2(f + 2N\hat{\xi})/N^2$ and can be used to construct confidence interval. This statistic is based on χ^2 of single marker-marker analysis, and readily extended to other statistics considered here. The statistic is implemented in program **HAPLO**. Ayres & Balding (2001) discussed MCMC method to obtain posterior distribution of LD statistic.

C programs for disease-marker analysis involving one or two biallelic markers are written to verify the gene counting as in **EH** and **EHPLUS** both for marker-marker case and disease-marker case. A C program **2LD** for obtaining two-locus disequibria and **fastEHPLUS** for obtaining permutation-based linkage disequilibrium measures are available from the section website. New algorithms developed for gene counting including handle of missing data were reported elsewhere.

Related works are Ceppellini et al. (1955), Hill (1975), Ott (1977), Xie & Ott (1993), Weir & Cockerham (1978, 1979), Weir (1990, 1996), Xie & Ott (1993), Excoffier & Slatkin (1995), Hawley & Kidd (1995), Long et al. (1995), Zaykin et al. (1995), Terwilliger & Ott (1994). However they either provide insufficient details or only deal with marker-marker analysis.

Appendix B

Computer programs

This appendix includes some subroutines and programs used in previous chapters.

B.1 C subroutines

Subroutine **fibonacci** is used to locate the interval containing optimal value in front of a unidimensional optimisation routine. It has been used in Chapter 4.

```
float fibonacci(float a,float b,float(*f)(float),float eps)
/*
Brian D Bunday & Gerald R Garside (1987). Optimisation Methods in Pascal.
Edward Arnold, finalized 11/11/98 JH Zhao
*/
{
float x1, x2, x3, x4, f2, f4, fib0, fib1, fib2;
int n, k;
n=15;
fib0 = 1;
fib1 = 1;
for (k = 2; k <= n; k++) {
  fib2 = fib1 + fib0;
  fib0 = fib1;
  fib1 = fib2;
}
```

```
x1 = a;
x3 = b;
if (n % 2) eps = -eps;
x2 = a + ((b - a) * fib0 + eps) / fib1;
f2 = (*f)(x2);
for (k = 2; k <= n; k++) {
  x4 = x1 - x2 + x3;
  f4 = (*f)(x4);
  if (f4 < f2) {
    if (x4 < x2) x3 = x2; else x1 = x2;
    x2 = x4;
    f2 = f4;
  } else if (x4 < x2) x1 = x4; else x3 = x4;
}
asave=x1;
bsave=x3;
return x2;
}
```

Subroutine **fmin** is popular for unidimensional optimisation (Forsythe et al. 1977; Press et al. 1992). An auxiliary routine *sign* is also given.

```
int fcount,maxn;
int sign(float a,float b)
/*an invent for Fortran SIGN(.) function in C*/
{
return((a>b)?1:((a==b)?0:-1));
}
float fmin(float ax, float bx, float (*f)(float), float tol)
/* This is Brent's method of univariate optimization adapted from
   Forsythe GE et al. (1977) and Press WH et al. (1992)
   JH Zhao 26/8/1998 IOP London
*/
```

```
{
int iter;
float a,b,d,fu,fv,fw,fx,p,q,r,tol1,tol2,u,v,w,x,xm;
float e=0.0,goldr=(3.0-sqrt(5.0))/2.0,maxiter=15,zeps=1.e-10;
a=ax;
b=bx;
w=v=x=bx;
fw=fv=fx=(*f)(x);
for (iter=1;iter<=maxiter;iter++) {
   xm=0.5*(a+b);
   tol2=2.0*(tol1=tol*fabs(x)+zeps);
   if (fabs(x-xm) <= (tol2-0.5*(b-a))) return x;
   if (fabs(e) > tol1) { /* fit parabola*/
      r=(x-w)*(fx-fv);
      q=(x-v)*(fx-fw);
      p=(x-v)*q-(x-w)*r;
      q=2.0*(q-r);
      if (q > 0.0) p = -p;
      q=fabs(q);
      r=e;
      e=d;
      if (fabs(p) < fabs(0.5*q*r) || p > q*(a-x) || p < q*(b-x))
      { /* parabolic interpolation step*/
         d=p/q;
         u=x+d;
         if (u-a < tol2 || b-u < tol2) d=sign(tol1,xm-x);
      } else /* golden section step*/
         d=goldr*(e=(x >= xm ? a-x : b-x));
   } else d=goldr*(e=(x >= xm ? a-x : b-x));
   u=(fabs(d) >= tol1 ? x+d : x+sign(tol1,d));
   fu=(*f)(u);
   if (fu <= fx) { /*update a,b,v,w,and x*/
```

```
      if (u >= x) a=x; else b=x;
      v = w; w = x; x = u;
      fv = fw; fw = fx; fx = fu;
   } else {
      if (u < x) a=u; else b=u;
      if (fu <= fw || w == x) {
         v=w; w=u;
         fv=fw; fw=fu;
      } else if (fu <= fv || v == x || v == w) {
        v=u; fv=fu;
      }
   }
}
/*perror("Maximum iterations exceeded in fmin");*/
return x;
}
```

Subroutine **digit**, due to Zhao & Sham (1998), has been used to sequentially list all possible patterns of mutations in Chapter 4.

```
int digit(int radix, int d[], int i)
/*digital addition*/
{
if (d[i] < radix) {
   ++d[i];
   goto ok;
   }
else {
   d[i] = 1;
   ++d[i + 1];
   if (d[i + 1] <= radix) goto ok;
   }
digit(radix, d, i+1);
```

```
ok:
return 0;
}
```

The binary search tree routines below are used in **fastEHPLUS**.

```
typedef struct node_type
{
  int key,n;
  struct node_type *left, *right;
} node;
node *itree(node*,int),*stree(node*,int),*dtree(node*,int);
void inorder(node*),preorder(node*),postorder(node*);
void rtree(node*),ptree(node*,int),etree(node*);
node *rt=NULL;
main()
{
int s[]={4,2,1,3,6,5,7,4,2,1,3,6,5,7},i;
for(i=0;i<14;i++) rt=itree(rt,s[i]);
printf("tree: root=%d\n\n",rt->key);
ptree(rt,0);
printf("\nPreorder:");
preorder(rt);
printf("\nInorder:");
inorder(rt);
printf("\nPostorder:");
postorder(rt);
etree(rt);
ptree(rt,0);
}
node *itree(node *r,int key)
/*insert and sort*/
{
```

```
  if (r==NULL)
  {
    r=(node*)malloc(sizeof(node));
    if (r==NULL)
    {
      printf("out of memory\n");exit(0);
    }
    r->left=r->right=NULL;
    r->key=key;
    r->n=1;
  }
  else
  if (key<r->key) r->left=itree(r->left,key);
  else if (key>r->key) r->right=itree(r->right,key);
  else r->n++;
  return r;
}
node *stree(node *t,int key)
/*search*/
{
  if (!t) return t;
  while (t->key!=key)
  {
    if (key<t->key) t=t->left;
    else t=t->right;
    if (t==0) break;
  }
  return t;
}
node *dtree(node *t,int key)
/*delete*/
{
```

```
node *p,*p2;

if (!t) return t;
if (t->key==key)
{
  if(t->left==t->right)
  {
    free(t);
    return NULL;
  }
  else if (t->left==NULL)
  {
    p=t->right;
    free(t);
    return p;
  }
  else if (t->right==NULL)
  {
    p=t->left;
    free(t);
    return p;
  }
  else
  {
    p2=t->right;
    p=t->right;
    while (p->left) p=p->left;
    p->left=t->left;
    free(t);
    return p2;
  }
}
```

```
  if (t->key<key) t->right=dtree(t->right,key);
  else t->left=dtree(t->left,key);
  return t;
}
void inorder(node *t)
/*left subtree->t->right subtree*/
{
  if (!t) return;
  inorder(t->left);
  printf("%d ",t->key);
  inorder(t->right);
}
void preorder(node *t)
/*t->left subtree->right subtree->t*/
{
  if (!t) return;
  printf("%d ",t->key);
  preorder(t->left);
  preorder(t->right);
}
void postorder(node *t)
/*left subtree->right subtree->t*/
{
  if (!t) return;
  postorder(t->left);
  postorder(t->right);
  printf("%d ",t->key);
}
void etree(node *t)
{
  printf("\nfree tree\n");
  rtree(t);
```

```
  rt=NULL;
}
void rtree(node *t)
/*left subtree->right subtree->t*/
{
  if (!t) return;
  rtree(t->left);
  rtree(t->right);
  printf("freeing %d\n",t->key);
  free(t);
}
void ptree(node *r,int l)
/*print tree inorder*/
{
  int i;
  if (!r) return;
  ptree(r->left,l+1);
  for (i=0;i<l;++i) printf("  ");
  printf("%d [%d]\n",r->key,r->n);
  ptree(r->right,l+1);
}
```

Subroutines **fbsize** and **pbsize** are used to obtain power of TDT and case-control designs in Chapters 4 and 7.

```
#include <stdio.h>
#include <math.h>
#define ERROR /*Science 273: 1516-17 13SEP1996*/
#undef ERROR  /*Science 275: 1327-30 28FEB1997*/
typedef struct normal{float m,v;} norm;
int fbsize(float,float),pbsize();
float n(norm,int);
float y,h,h1,h2,p_A,lambda_o,lambda_s,n1,n2,n3;
```

```
float models[12][2]={
    4.0, 0.01,      4.0, 0.10,      4.0, 0.50,      4.0, 0.80,
    2.0, 0.01,      2.0, 0.10,      2.0, 0.50,      2.0, 0.80,
    1.5, 0.01,      1.5, 0.10,      1.5, 0.50,      1.5, 0.80};
main()
{
float gamma,p;
int i;
/*Risch & Merikangas 1996*/
printf("\nThe family-based result: \n");
printf("\ngamma   p      Y     N_asp    P_A    Het
   N_tdt   Het N_asp/tdt  L_o   L_s\n\n");
for(i=0;i<12;i++) {
  gamma=models[i][0];
  p=models[i][1];
  fbsize(gamma,p);
  if((i+1)%4==0) printf("\n");
}
/*APOE-4, Scott WK, Pericak-Vance, MA & Haines JL*/
/*Genetic analysis of complex diseases 1327*/
gamma=4.5;
p=0.15;
printf("Alzheimer's:\n\n");
fbsize(gamma,p);
/*Long AD, Grote MN & Langley CH 1328*/
pbsize();
}

int fbsize(float gamma,float p)
/*Jing Hua Zhao 30-12-98*/
{
norm nl, al, na, aa;
```

```
float k,va,vd,q,w;
q=1-p;
k=pow(p*gamma+q,2);
va=2*p*q*pow((gamma-1)*(p*gamma+q),2);
vd=pow(p*q*pow(gamma-1,2),2);
#ifdef DEBUG
printf("K=%f VA=%f VD=%f\n",k,va,vd);
#endif
w=p*q*pow(gamma-1,2)/pow(p*gamma+q,2);
y=(1+w)/(2+w);
lambda_s=pow(1+0.5*w,2);
lambda_o=1+w;
h=h1=p*q*(gamma+1)/(p*gamma+q);
p_A=gamma/(gamma+1);
/*ASP*/
nl.m=0;
nl.v=1;
al.m=2*y-1;
#ifdef ERROR
al.v=0;
#else
al.v=4*y*(1-y);
#endif
n1=n(al,1);
/*TDT*/
na.m=0;
na.v=1;
aa.m=sqrt(h)*(gamma-1)/(gamma+1);
aa.v=1-h*pow((gamma-1)/(gamma+1),2);
n2=n(aa,2);
/*ASP-TDT*/
h=h2=p*q*pow(gamma+1,2)/(2*pow(p*gamma+q,2)+p*q*pow(gamma-1,2));
```

```
aa.m=sqrt(h)*(gamma-1)/(gamma+1);
aa.v=1-h*pow((gamma-1)/(gamma+1),2);
n3=n(aa,3);
printf("%4.1f%5.2f|%6.3f%10.0f|%6.3f|%6.3f%8.0f%6.3f%8.0f|%6.2f%6.2f\n",
       gamma,p,y,n1,p_A,h1,n2,h2,n3,lambda_o,lambda_s);
return 0;
}

float n(norm all,int op)
/*m=0,v=1 under the null hypotheses*/
{
float z_alpha,z1_beta;
float s,m,v;
m=all.m;
v=all.v;
z1_beta=-0.84;                      /*1-beta=0.8,-.84162123*/
switch(op) {
case 1:z_alpha=3.72;break;          /*alpha=1E-4,3.7190165*/
default:z_alpha=5.33;break;         /*alpha=5E-8,5.3267239*/
}
s=pow((z_alpha-sqrt(v)*z1_beta)/m,2)/2; /*shared/transmitted for each parent*/
if(op==3) s/=2;                        /*the sample size is halved*/
return (s);
}

#define x2_alpha 29.72 /*Q=29.7168*/
#define z_alpha 5.45   /*5.4513104*/

int pbsize()
{
int i,j;
float z1_beta,gamma,lambda,pi,p,q,n,k,s,kp[3]={0.01,0.05,0.10};
```

```
/*
\alpha=5e-8,\beta=0.8, \alpha would give 5\% genome-wide significance level
\lambda is the NCP from the marginal table
\pi is the pr(Affected|aa)
*/
gamma=4.5; /*again for Alzheimer's*/
p=0.15;
q=1-p;
pi=0.065;  /*0.07 generates 163, equivalent to ASP*/
z1_beta=-0.84;
k=pi*pow(gamma*p+q,2);
s=(1-pi*pow(gamma,2))*pow(p,2)+(1-pi*gamma)*2*p*q+(1-pi)*pow(q,2);
/*LGL formula*/
lambda=pi*(pow(gamma,2)*p+q-pow(gamma*p+q,2))/(1-pi*pow(gamma*p+q,2));
/*my own*/
lambda=pi*p*q*pow(gamma-1,2)/(1-pi*pow(gamma*p+q,2));
/*not sure about +/-!*/
n=pow(z1_beta+z_alpha,2)/lambda;
/*may be used to correct for population prevalence*/
printf("\nThe population-based result: Kp=%f Kq=%f n=%10.0f\n",k,s,n);
printf("\nRandom ascertainment with disease prevalence\n");
printf("\n         1%%         5%%        10%%\n\n");
for(i=0;i<12;i++) {
  gamma=models[i][0];
  p=models[i][1];
  q=1-p;
  for(j=0;j<3;j++) {
    k=kp[j];
    pi=k/pow(gamma*p+q,2);
    lambda=pi*p*q*pow(gamma-1,2)/(1-pi*pow(gamma*p+q,2));
    n=pow(z1_beta-z_alpha,2)/lambda;
    printf(" %10.f",n);
```

```
  } printf("\n");
  if((i+1)%4==0) printf("\n");
}
printf("This is only an approximation, a more accurate result\n");
printf("can be obtained by Fisher's exact test\n");
return 0;
}
```

B.2 SAS programs

Program **power** obtains noncentrality parameter of χ^2_f distribution, $f = 1, \ldots, 30$, for a given power and significant level α.

```
title;
data power;
do f=1 to 30;
   alpha=0.05;
   power=0.8;
   c=cinv(1-alpha,f);
   NCP=cnonct(c,f,1-power);
   p=probchi(c,f,c);
   d=cinv(1-alpha,f,0.87);
   output;
end;
proc print;run;
```

Program **ncp** includes a macro to obtain the required sample sizes using means of log-likelihood ratio tests.

```
/*Adapted from ncp.sas (use macro) JH Zhao 1-8-2002, redo 8-8-2002*/
options nocenter nonumber;
%macro n(data=abc);
data abc;
  set &data;
```

```
  array m{7} T H F R D P G;
  array s{7} st sh sf sr sd sp sg;
  array nn{7} nnt nnh nnf nnr nnd nnp nng;
  do i=1 to 7;
     if i=3 then nn[i]=ceil(18.3/(m[i]/nobs));
     else nn[i]=ceil(17.4/(m[i]/nobs));
  end;
  array n{7} nt nh nf nr nd np ng;
  array k{7} kt kh kf kr kd kp kg;
  do i=1 to 7;
     if (i=3) then do;
        n[i]=ceil(18.3/(m[i]-7)*nobs);
        k[i]=ceil(18.3/((s[i]*s[i]-2*7)/4)*nobs);
     end;
     else do;
        n[i]=ceil(17.4/(m[i]-6)*nobs);
        k[i]=ceil(17.4/((s[i]*s[i]-2*6)/4)*nobs);
     end;
  end;
  drop i;
run;
proc print noobs;
     var model nnt nnr nnd nnf nnh nnp nng;
     where model ne 'NULL';
run;
proc print noobs;
     var model nt nr nd nf nh np ng;
     where model ne 'NULL';
run;
proc print noobs;
     var model kt kr kd kf kh kp kg;
     where model ne 'NULL';
```

```
run;
proc print noobs;
     var model t r d f h p g st sr sd sf sh sp sg;
     format _numeric_ 7.2;
run;
%mend n;
data abc10000;
  nobs=10000;
  do model='NULL','RR','RD','CR','CD','ALZ';
     input name$ T H F R D P G/name$ st sh sf sr sd sp sg;
     output;
  end;
cards;
mean     6.086     6.161     6.244     6.153     6.159     6.144     6.157
 std     3.481     3.552     3.595     3.550     3.557     3.533     3.548
mean  3199.286  3199.277  3199.285  3199.285  2356.602  3063.708  3197.999
 std   111.160   111.121   111.105   111.105    77.514   104.428   110.949
mean  1041.946   907.501  1041.946   907.501  1041.946   895.769   907.543
 std    57.352    52.210    57.357    52.202    57.352    50.888    52.217
mean   804.946   716.690   805.366   716.718   541.284   710.195   716.737
 std    60.346    57.130    60.390    57.128    42.141    56.103    57.127
mean   261.983   252.311   262.728   252.311   262.585   251.125   252.346
 std    31.544    30.534    31.622    30.536    31.631    30.246    30.529
mean   268.792   270.229   270.977   270.353   261.748   269.156   270.263
 std    32.161    32.490    32.454    32.487    31.324    32.233    32.499
;
data abc1000;
  nobs=1000;
  do model='NULL','RR','RD','CR','CD','ALZ';
     input name$ T H F R D P G/name$ st sh sf sr sd sp sg;
     output;
  end;
```

```
cards;
mean     5.393     6.173     6.377     6.148     6.084     6.032     6.177
 std     2.802     3.366     3.416     3.330     3.219     3.221     3.379
mean   327.218   327.218   327.295   327.218   242.027   312.990   327.172
 std    33.957    33.961    33.977    33.959    23.160    31.830    33.919
mean   109.637    96.629   109.785    96.629   109.637    95.050    96.638
 std    19.155    18.022    19.231    18.023    19.155    17.463    18.025
mean    85.210    77.086    86.309    77.089    59.308    76.035    77.082
 std    18.618    17.870    18.807    17.870    13.341    17.397    17.872
mean    30.568    30.370    32.022    30.370    31.399    30.045    30.370
 std    10.170    10.204    10.425    10.205    10.315    10.037    10.204
mean    31.881    32.896    33.842    32.881    31.944    32.560    32.901
 std    10.876    11.132    11.300    11.138    10.725    10.980    11.138
;
%n(data=abc10000);
%n(data=abc1000);
```

Macro **setf** obtains disease models compatible with population disease prevalence and genotypic distribution. An example of Alzheimer's disease is also provided.

```
/*********************************************************/
/*This SAS macro gives a set of kp compatible q and f's.   */
/*      kp=population prevalence                           */
/*f0,f1,f2=the penetrance                                  */
/*       q=the putative disease allele frequency           */
/*Corrected 4/8/97 IOP JH ZHAO                             */
/*********************************************************/
%macro setf(setq=,setf0=,setf1=,setf2=);
title;
options nocenter nodate nonumber;
proc iml;
q=&setq;p=1-q;
f0=&setf0;f1=&setf1;f2=&setf2;
```

```
p1=p;p2=q;
k=f0*p**2+f1*2*p*q+f2*q**2;
f=(f0//f1//f2);
b=(f0*p**2/k)//(f1*2*p*q/k)//(f2*q**2/k);
print k q f b;
step1=q/10;
step2=p*q/10;
do q=0.0 to 1.0 by step1;
   p=1-q;
   do d=-p*q to p*q by step2;
      h11=p*p1+d;
      h12=p*p2-d;
      h21=q*p1-d;
      h22=q*p2+d;
      h=h11+h12+h21+h22;
      a=(  (h11*h11 || 2*h11*h21           || h21*h21)//
         (2*h11*h12 || 2*h11*h22+2*h12*h21 || 2*h21*h22)//
           (h12*h12 || 2*h12*h22           || h22*h22)
        )/k;
      f=ginv(a)*b;
      kp=f[1]*p**2+f[2]*2*p*q+f[3]*q**2;
      hh=h11//h12//h21//h22;
      c=a[+,]*k;
      if (h11>=0 && h12>=0 && h21>=0 && h22>=0) &&
         (f[1]>=0 && f[1] <= f[2] && f[2]<=f[3] && f[3]<=1)
      then print hh q d a f kp;
   end;
end;
quit;
%mend setf;
%setf(setq=0.13,setf0=0.05,setf1=0.2,setf2=0.8);
```

B.3 QBASIC program

N.BAS illustrates how mixture of χ^2 is used for obtaining required sample size and has been used in Chapter 6.

```
DECLARE SUB ssize (pd, p1, r!, g!)
PRINT
PRINT "pd", "p1", "r", "g", "n (t=10cM)"
PRINT
CALL ssize(.1, .1, 4, 15): CALL ssize(.5, .1, 4, 15)
CALL ssize(.1, .3, 4, 15): CALL ssize(.5, .3, 4, 15)
CALL ssize(.1, .5, 4, 15): CALL ssize(.5, .5, 4, 15)
CALL ssize(.1, .7, 4, 15): CALL ssize(.5, .7, 4, 15)
CALL ssize(.1, .9, 4, 15): CALL ssize(.5, .9, 4, 15)
PRINT
CALL ssize(.1, .1, 4, 50): CALL ssize(.5, .1, 4, 50)
CALL ssize(.1, .3, 4, 50): CALL ssize(.5, .3, 4, 50)
CALL ssize(.1, .5, 4, 50): CALL ssize(.5, .5, 4, 50)
CALL ssize(.1, .7, 4, 50): CALL ssize(.5, .7, 4, 50)
CALL ssize(.1, .9, 4, 50): CALL ssize(.5, .9, 4, 50)
END

' sample size required for LLD with S families
'
' pd=disease allele frequency
' pn=normal allele frequency
' p1=associated allele frequency for biallelic locus
' n=sample size
' t=theta, recombination fraction, 10cM and Haldane function
' r=relative risk assuming multiplicative model
' psi as in Haseman and Elston (1972)
' K=population prevalence
' g=generations
```

```
' s=required sample size
' critical value for 0.5*X^2(1)+0.5*X^2(2)=3.77*2*2.3025851
' x2n=noncentrality parameter for alpha=0.0001, beta=0.2
' di npnchi(2,3.77*2*2.3025851,0.2) " " nchi(2,23.978829,3.77*2*2.3025851)
' 23.978829 .1999992
'
SUB ssize (pd, p1, r, g)
t = .0906  '10cM
psi = t ^ 2 + (1 - t) ^ 2
pn = 1 - pd
k = (pd * r + pn) ^ 2
ncp = (4 * psi / k) * ((r - 1)*pd) ^ 2 * ((1 - p1) / p1) * EXP(-2 * t * g)
x2n = 23.978829#
n = x2n / ncp
PRINT pd, p1, r, g, n

END SUB
```

Bibliography

Abecasis, G. R., L. R. Cardon, and O. C. Cookson (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet 66*, 279–292.

Abecasis, G. R., W. O. C. Cookson, and L. R. Cardon (2000). Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics 8*, 545–551.

Abel, L. and B. Müller-Myhsok (1998). Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. *Am J Hum Genet 63*, 664–667.

Abreu, P. C., D. A. Greenberg, and S. E. Hodge (1999). Direct power comparisons between simple lod scores and npl scores for linkage analysis in complex diseases. *Am J Hum Genet 65*, 847–857.

Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet 60*(3), 676–690.

Allison, D. B., M. Heo, N. Kaplan, and E. R. Martin (1999). Sibling-based tests of linkage and association for quantitative trait. *Am J Hum Genet 64*, 1754–1764.

Almasy, L. and J. Blangero (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet 62*, 1198–1211.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet 54*, 535–543.

Amos, C. I. and M. de Andrade (2001). Genetic linkage methods for quantitative traits. *Stat Meth Med Res 10*, 3–25.

Amos, C. I., M. de Andrade, and D. K. Zhu (2001). Comparison of multivariate tests for genetic linkage. *Hum Hered 51*, 133–144.

Amos, C. I., D. K. Zhu, and E. Boerwinkle (1996). Assessing genetic linkage and association with robust components of variance approaches. *Ann Hum Genet 60*, 143–160.

Ayres, K. L. and D. J. Balding (2001). Measuring gametic disequilibrium from multilocus data. *Genetics 17*, 413–423.

Bacanu, S.-A., B. Devlin, and K. Roeder (2000). The power of genomic control. *Am J Hum Genet 66*, 1933–1944.

Badner, J. A., E. S. Gershon, and L. R. Goldin (1999). Reply to baron. *Am J Hum Genet 64*, 1246–1248.

Badner, J. A., E. S. Gershon, and G. L. R. (1998). Optimal ascertainment strategies to detect linkage to common disease alleles. *Am J Hum Genet 63*, 880–888.

Bahlo, M. and R. C. Griffiths (2000). Inference from gene trees in a subdivided population. *Theor Popul Biol 57*, 79–95.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions on Markov processes. *Inequalities 3*, 1–8.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization techniques occurring in the statistical analysis of probablistic functions of Markov chains. *Ann Math Stat 41*, 164–171.

Baur, M. P. and M. Knapp (1997). Association studies in genet epidemiol. In I.-H. Pawlowitzki, J. H. Edwards, and E. A. Thompson (Eds.), *Genetic Mapping of Disease Genes*, pp. 159–172. London: Acadmic Press.

Beaty, T. H. (1997). Evolving methods in genet epidemiol: I. analysis of genetic and environmental factors in family studies. *Epidemiol Rev 19*, 14–23.

Bengtsson, B. O. and G. Thomson (1981). Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens 18*, 356–363.

Bennett, J. H. (1954). On the theory of random mating. *Annals of Eugenics 18*, 311–317.

Bickeböller, H. and F. Clerget-Darpoux (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol 12*, 865–870.

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

Boehnke, M. (1986). Estimating the power of a proposed linkage study: a practical computer simulation approach. *Am J Hum Genet 39*, 513–527.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: implication for the positional cloning of human disease genes. *Am J Hum Genet 55*, 379–390.

Boehnke, M. and C. D. Langefeld (1998). Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am J Hum Genet 62*, 950–961.

Bonney, G. E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: regressive models. *Am J Med Genet 18*, 731–749.

Bonney, G. E. (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics 42*, 611–625.

Borecki, I. B., M. A. Province, and D. C. Rao (1994). Power of segregation analysis for detection of major gene effects on quantitative traits. *Genet Epidemiol 11*, 409–418.

Borecki, I. B., M. A. Province, and D. C. Rao (1995). Inferring a major gene for quantitative traits by using segregation analysis with tests on transmission probabilities: how often do we miss ? *Am J Hum Genet 56*, 319–326.

Boyle, C. R. and R. C. Elston (1979). Multifactorial genetic models for quantitative traits in humans. *Biometrics 35*, 55–68.

Broman, K. W. and J. L. Weber (2000). Characterization of human crossover interference. *Am J Hum Genet 66*, 1911–1926.

Brown, P. O. and D. Botstein (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet 21*, 33–37.

Camp, N. J. (1997). Genomewide transmission/disequilibrium testing-consideration of the genotypic relative risks at disease loci. *Am J Hum Genet 61*, 1424–1430.

Camp, N. J. (1999). Genomewide transmission/disequilibrium testing: a correction. *Am J Hum Genet 64*, 1485–1487.

Cantor, C. (1994). Can computational science keep up with evolving technology for genome mapping and sequences? In S. Suhai (Ed.), *Computational Methods in Genome Research*, pp. 1–18. New York: Plenum Press.

Cardon, L. R. and J. I. Bell (2001). Association study designs for complex diseases. *Nature Reviews Genetics 2*, 91–99.

Casella, G. and R. L. Berger (1990). *Statistical Inference*. Belmont, California: Wadsworth Inc.

Cavalli-Sforza, L. and W. F. Bodmer (1971). *The Genetics of Human Populations*. San Francisco: W. H. Freeman.

Ceppellini, R., M. Siniscalco, and C. A. B. Smith (1955). The estimation of gene frequencies in a random mating population. *Ann Hum Genet 20*, 97–115.

Cervnio, A. C. L. and A. V. S. Hill (2000). Comparison of tests for association and linkage in incomplete families. *Am J Hum Genet 67*, 120–132.

Chakraborty, R. and K. M. Weiss (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA 85*, 9119–9123.

Chapman, D. G. and R. C. Meng (1966). The power of chi-square tests for contingency tables. *J Amer Stat Assoc 61*, 965–975.

Chapman, N. H. and E. M. Wijisman (1998). Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am J Hum Genet 63*, 1872–1885.

Chernoff, H. and E. Lander (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J Stat Plan Infer 43*, 19–40.

Chiano, M. N. and D. G. Clayton (1998). Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet 62*, 55–60.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan, J. Stengård, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C. F. Sing (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet 63*, 595–612.

Clayton, D. (1999). A generalization of the transmission/disequilbrium test for uncertain-haplotype transmission. *Am J Hum Genet 65*, 1170–1177.

Clayton, D. (2000). Linkage disequilibrium mapping of disease susceptibility genes in human populations. *International Statistical Review 68*, 23–43.

Clayton, D. and H. Jones (1999). Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet 65*, 1161–1169.

Clerget-Darpoux, F., C. Bonaïti-Pellié, and J. Hochez (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics 42*, 393–399.

Cleves, M. A., J. M. Olson, and K. B. Jacobs (1997). Exact transmission/disequilibrium tests with multiallelic markers. *Genet Epidemiol 14* (4), 337–347.

Collins, A., C. Lonjou, and N. E. Morton (1999). Genet epidemiol of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA 96*, 15173–15177.

Cooper, G., W. Amos, D. Hoffman, and D. C. Rubinson (1996). Network analysis of human Y microsatelliete haplotypes. *Hum Mol Genet 5*, 1759–1766.

Cottingham, R. W. J., R. M. Idury, and A. A. Schaffer (1993). Faster sequential genetic linkage computations. *Am J Hum Genet 53*, 252–263.

Cox, A., N. J. Camp, M. J. H. Nicklin, F. S. di Giovine, and G. W. Duff (1998). An analysis of linkage disequilibrium in the interleukin-1 gene cluster, using a novel grouping method for multiallelic markers. *Am J Hum Genet 62*, 1180–1188.

Crow, J. F. (2000). Two centuries of genetics: a view from halftime. *Annual Review of Genomics and Human Genetics 1*, 21–40.

Curtis, D. (1997). Use of siblings as controls in case-control association studies. *Ann Hum Genet 61*, 319–333.

Curtis, D. and P. C. Sham (1995a). Model-free linkage analysis using likelihoods. *Am J Hum Genet 57*, 703–716.

Curtis, D. and P. C. Sham (1995b). A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet 56*, 811–812.

Darvasi, A. (1998). Experimental strategies for the genetic dissection of complex traits in animal models. *Nat Genet 18*, 19–24.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 64*, 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 74*, 33–43.

Davis, S., L. R. Goldin, and D. E. Weeks (1997). SimIBD: a powerful robust nonparametric method for detecting linkage in general pedigrees. In I.-H. Pawlowitziki, J. H. Edwards, and E. A. Thompson (Eds.), *Genetic Mapping of Disease Genes*, pp. 189–204. London: Academic Press.

Davis, S., M. Schroeder, L. R. Goldin, and D. E. Weeks (1996). Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *Am J Hum Genet 58*, 867–880.

Davis, S. and D. E. Weeks (1997). Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *Am J Hum Genet 61*, 1431–1444.

de la Chapelle, A. (1993). Disease gene mapping in isolated human populations: the example of finland. *Journal of Medical Genetics 30*, 857–865.

de La Chapelle, A. and F. A. Wright (1998). Linkage disequilibrium mapping in isolated populations: the examples of Finland revisited. *Proc Natl Acad Sci 95*, 12416–12423.

Dempster, A. P., N. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B 39*, 1–38.

Devlin, B. and K. Roeder (1999). Genomic control for association studies. *Biometrics 55*, 997–1004.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature 380*, 152–154.

Doerge, R. W., Z.-B. Zeng, and B. Weir (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science 12*(3), 195–219.

Dudbridge, F., P. C. Koeleman, and J. A. Todd (2000). Unbiased application of the tranmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet 66*, 2009–2012.

Dudoit, S. and T. P. Speed (2000). A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. *Biostatistics 1*, 1–26.

Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent (1999). Expression profiling using cDNA microarrrays. *Nat Genet 21*, 10–14.

Dupuis, J. and P. van Eerdewegh (2000). Multipoint linkage analysis of the pseudoautosomal regions, using affected sibling pairs. *Am J Hum Genet 67*, 462–475.

Durner, M., D. A. Greenberg, and S. E. Hodge (1992). Inter- and intrafamilial heterogeneity: effective sampling strategies and comparison of analysis methods. *Am J Hum Genet 51*, 859–870.

Durner, M., V. J. Vieland, and D. A. Greenberg (1999). Further evidence for the increased power of lod scores compared with nonparametric methods. *Am J Hum Genet 94*(1), 281–289.

Elandt-Johnson, R. C. (1970). Segregation analysis for complex modes of inheritance. *Am J Hum Genet 22*, 129–144.

Elandt-Johnson, R. C. (1971). *Probability Models and Statistical Methods in Genetics*. John Wiley & Sons, Inc.

Elston, R. C. (1980). segregation analysis. In J. H. Mielke and M. H. Crawford (Eds.), *Current Developments in Anthropological Genetics*, Volume I., pp. 327–354. New York: Plenum Press.

Elston, R. C. (1992). P values, power, and pitfalls in the linkage analysis of psychiatric disorders. In E. S. Gershon and C. R. Cloninger (Eds.), *Genetic Approaches to Mental Disorders*, pp. 3–21. Washington, D.C.: American Psychiatric Press.

Elston, R. C. (2000). Introduction and overview. *Stat Meth Med Res 9*, 527–541.

Elston, R. C., S. Buxbaum, K. B. Jacobs, and J. M. Olson (2000). Haseman and elston revisited. *Genet Epidemiol 19*, 1–17.

Elston, R. C. and J. Stewart (1971). A general model for the genetic analysis of pedigree data. *Hum Hered 21*, 523–542.

Ewens, W. J. (1982). Aspects of parameter estimation in ascertainment sampling schemes. *Am J Hum Genet 34*, 853–865.

Ewens, W. J. (1991). Ascertainment bias and their resolution in biological surveys. In C. R. Rao and R. Chakraborty (Eds.), *Handbook of Statistics*, Volume 8, pp. 29–61. Elsevier Science Publisher.

Eyre-Walker, A. and P. D. Keightley (1999). High genomic deleterious mutation rates in hominids. *Nature 397*, 344–347.

Falk, C. T. and P. Rubinstein (1987). Haplotype relative risk: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet 51*, 227–233.

Fallin, D., A. Cohen, L. Essioux, I. Chumakov, M. Blumenfeld, D. Cohen, and N. J. Schork (2001). Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res 11*, 143–151.

Farrall, M. and D. E. Weeks (1998). Mutational mechanisms for generating microsatellite allele-frequency distributions: an analysis of 4,558 markers. *Am J Hum Genet 62*, 1260–1262.

Feder, J. N., A. Gnirke, W. Thomas, Z. Tsuchihashi, D. A. Ruddy, A. Basava, F. Dormishian, J. Domingo, R., M. C. Ellis, A. Fullan, L. M. Hinton, N. L. Jones, B. E. Kimmel, G. S. Kronmal, P. Lauer, V. K. Lee, D. B. Loeb, F. A. Mapa, E. McClelland, N. C. Meyer, G. A. Mintier, N. Moeller, T. Moore, E. Morikang, C. E. Prass, L. Quintana, S. M. Starnes, R. C. Schatzman, K. J. Brunke, D. T. Drayna, N. J. Risch, B. R. Bacon, and R. K. Wolff (1996). A novel mhc class i-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet 13*, 399–408.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution 17*, 368–376.

Felsenstein, J., M. K. Kuhner, J. Yamato, and P. Beerli (1999). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. *IMS Lecture Notes Monographic Series 33*, 163–185.

Forsythe, G. E., M. A. Malcolm, and C. B. Moler (1977). *Computer Methods for Mathematical Computations*. Englewood Cliffs, N.J.: Printice-Hall.

Fortna, A. and K. Gardiner (2001). Genomic sequence analysis tools: a user's guide. *Trends in Genetics 17*, 158–164.

Fu, Y.-X. and W.-H. Li (1999). Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor Popul Biol 56*, 1–10.

Fulker, D. W. and S. S. Cherny (1996). An improved multipoint sib-pair analysis of quantitative traits. *Behav Genet 26*, 527–532.

Fulker, D. W., S. S. Cherny, and L. R. Cardon (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am J Hum Genet 56*, 1224–1233.

Fulker, D. W., S. S. Cherny, P. C. Sham, and J. K. Hewitt (1999). Combined linkage and

association sib-pair analysis for quantitative traits. *Am J Hum Genet 64* (1), 259–267.

Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. *Ann Math Stat 15*, 25–57.

George, V., H. K. Tiwari, X. Zhu, and R. C. Elston (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet 65*, 236–245.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Ann Statist 22*, 1993–2010.

Ghosh, S. and P. P. Majumder (2000). Mapping a quantitative trait locus via the em algorithm and bayesian classification. *Genet Epidemiol 19*, 97–126.

Gillespie, J. H. (1998). *Population Genetics: A Concise Guide*. Baltimore and London: The Johns Hopkins University Press.

Goldgar, D. E. (1990). Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet 47*, 957–967.

Goldgar, D. E. and R. O. Oniki (1992). Comparison of a multipoint identity by descent method with parametric multipoint linkage analysis for mapping quantitative traits. *Am J Hum Genet 50*, 598–606.

Göring, H. H. H. and J. D. Terwilliger (2000). Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singleton when the mode of inheritance cannot be accurately specified. *Am J Hum Genet 66*, 1310–1327.

Graham, J., J. Curran, and B. S. Weir (2000). Conditional genotypic probabilities for microsatellite loci. *Genetics 158*, 1973–1980.

Graham, J. and E. A. Thompson (1998). Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am J Hum Genet 63*, 1517–1530.

Graham, J. and E. A. Thompson (2000). A coalescent model of ancestry for a rare allele. *Genetics 156*, 375–384.

Greenberg, D. A. (1989). Inferring mode of inheritance by comparison of lod scores. *Am J Med Genet 34*, 480–486.

Greenberg, D. A., P. Abreu, and S. E. Hodge (1998). The power to detect linkage in complex disease by means of simple lod-score analyses. *Am J Hum Genet 63*, 870–879.

Griffiths, R. C. and P. Marjoram (1996). Ancestral inference from samples of DNA sequences with recombination. *J Comp Biol 3*, 479–502.

Griffiths, R. C. and S. Tavaré (1994a). Ancestral inference in population genetics. *Statistical Science 9*, 307–319.

Griffiths, R. C. and S. Tavaré (1994b). Simulatin probability distributions in the coalescent. *Theor Popul Biol 46*, 131–159.

Grimmett, G. R. and D. R. Stirzaker (1992). *Probability and Random Processes* (Second ed.). Oxford: Clarendon Press.

Gudbjartsson, D. F., K. Jonasson, M. L. Frigge, and A. Kong (2000). Allegro, a new computer program for multipoint linkage analysis. *Nat Genet 25*, 12–13.

Guenther, W. C. (1977). Power and sample size for approximate chi-square tests. *The American Statistician 31*, 83–85.

Guo, S.-W. (2000a). Gene-environmental interactions and the affected-sib-pair designs. *Hum Hered 50*, 271–285.

Guo, S.-W. (2000b). Genetic mapping of complex traits: promises, problems, and prospects. *Theor Popul Biol 57*, 1–11.

Guo, S.-W. and E. A. Thompson (1992a). A Monte Carlo method for combined segregation and linkage analysis. *Am J Hum Genet 51*, 1111–1126.

Guo, S. W. and E. A. Thompson (1992b). Performing the exact test of hardy-weinberg proportion for multiple alleles. *Biometrics 48*, 361–372.

Guo, S.-W. and M. Xiong (1997). Estimating the age of mutant disease alleles based on linkage disequilbrium. *Hum Hered 47*, 315–337.

Haldane, J. B. S. and C. A. B. Smith (1947). A new estimate of the linkage between genes for haemophilia and colour-blindness in man. *Annals of Eugenics 14*, 10–31.

Hampe, J., T. Wienker, S. Schreiber, and P. Nurnberg (1998). POPSIM: a general population program. *Bioinformatics 14*, 458–464.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox, J. A. Schneider, D. S. Moulin, and J. B. Clegg (1997). Archaic Arican and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet 60*, 772–789.

Hartl, D. L. (2000). *A primer of Population Genetics* (Third ed.). Sunderland, Massachusetts: Sinauer Associates, Inc.

Hartl, D. L. and A. G. Clark (1997). *Principles of Population Genetics* (Third ed.). Sunderland, Massachusetts: Sinauer Associates, Inc.

Haseman, J. K. and R. C. Elston (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet 2*(1), 3–19.

Hasstedt, S. and P. Cartwright (1981). PAP: Pedigree analysis package. Technical Report 13, Department of Medical Computing and Biophysics, University of Utah.

Hasstedt, S. J. (1982). A mixed-model likelihood approximation on large pedigrees. *Comput. Biomed. Res. 15*, 295–307.

Hasstedt, S. J. (1994). Pedigree analysis package, University of Utah.

Hasstedt, S. J., M. Hoffman, M. F. Leppert, and S. C. Elbein (1997). Recessive inheritance of obesity in familial non-insulin-dependent diabetes mellitus, and lack of linkage to nine candidate genes. *Am J Hum Genet 61*, 668–677.

Hasstedt, S. J., D. A. Meyers, and D. G. Marsh (1983). Inheritance of immunoglobin E: genetic model fitting. *Am. J. Med. Genet. 14*, 61–66.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Hauser, E. R. and M. Boehnke (1998). Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics 54*, 1238–1246.

Hauser, E. R., M. Boehnke, S.-W. Guo, and N. Risch (1996). Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. *Genet Epidemiol 13*, 117–137.

Hawley, M. E. and K. K. Kidd (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered 86*, 409–411.

Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet 61*, 748–760.

Hill, W. G. (1974). Disequilibrium among several linked neutral genes in finite population. II. variances and covariances of disequilibra. *Theor Popul Biol 6*, 184–198.

Hill, W. G. (1975). Tests for association of gene frequencies at several loci in random mating diploid populations. *Biometrics 31*, 881–888.

Hill, W. G. and A. Robertson (1968). Linkage disequilibrium in finite populations. *Theor Popul Biol 38*, 226–231.

Hinds, D. and N. Risch (1996). The aspex package: affected sib-pair mapping. ftp://lahmed.stanford.edu/pub/aspex/index.html.

Hirotsu, C., S. Aoki, T. Inada, and Y. Kitao (2001). An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis. *Biometrics 57*, 769–778.

Hodge, S. E. (1998). Exact elods and exact power for affected sib pairs analyzed for linkage under simple right and wrong models. *Am J Med Genet 81*, 66–72.

Hodge, S. E., P. C. Abreu, and D. A. Greenberg (1997). Magnitude of type I error when single-locus linkage analysis is maximized over models: a simulation study. *Am J Hum Genet 60*, 217–227.

Hodge, S. E. and R. C. Elston (1994). Lods wrods mods: The interpretation of lod scores calculated under different models. *Genetic Epidemiol 11*, 329–342.

Hogenesch, J. B., K. A. Ching, S. Batalov, A. I. Su, J. R. Walker, Y. Zhou, S. A. Kay, P. G. Schultz, and M. P. Cooke (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell 106*, 413–415.

Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet 52*, 362–374.

Hopper, J. L. and J. D. Mathews (1982). Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet 46*, 373–383.

Horvath, S. and N. M. Laird (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet 63*, 1886–1897.

Horvath, S., C. Windemuth, and M. Knapp (2000). The disequilibrium maximum-likelihood-binomial test does not replace the transmission/disequilibrium test. *Am J Hum Genet 67*, 531–534.

Huang, J. and Y. Jiang (2001). The score statistic of the LD-Lod analysis: detecting linkage adaptive to linkage disequilibrium. *Hum Hered 52*, 83–98.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology 7*, 1–44.

Idury, R. M. and R. C. Elston (1997). A faster and more general hidden Markov model algorithms for multipoint likelihood calculations. *Hum Hered 47*, 197–202.

Jacquard, A. (1974). *The Genetic Structure of Populations*. New York: Springer-Verlag.

Johnson, G. C. L. and J. A. Todd (2000). Strategies in complex disease mapping. *Current Opinion in Genetics and Development 10*, 330–334.

Jorde, L. B. (1995). Invited editorial: linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet 56*, 11–14.

Kaplan, N. L., W. G. Hill, and B. S. Weir (1995). Likelihood methods for locating disease genes in nonequilibrium populations. *Am J Hum Genet 56*, 18–32.

Kaplan, N. L., E. R. Martin, and B. S. Weir (1997). Power studies for the transmission/disequilibrium tests with multiple alleles. *Am J Hum Genet 60*(3), 691–702.

Keavney, B., C. A. McKenzie, J. M. C. Connell, C. Julier, P. J. Ratcliffe, E. Sobel, M. Lathrop, and M. Farrall (1998). Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum Mol Genet 11*, 1745–1751.

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

Kidd, J. R., A. J. Pakstis, H. Zhao, R.-B. Lu, F. E. Okonofua, A. Odunsi, E. Grigorenko, B. Bonne-Tamir, J. Friedlaender, L. O. Schulz, J. Parnas, and K. K. Kidd (2000). Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, pah, in a global representation of population. *Am J Hum Genet 66*, 1882–1899.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Applications 13*, 235–248.

Kingman, J. F. C. (2000). Origins of the coalescent: 1974-1982. *Genetics 156*, 1461–1463.

Klitz, W., J. Stephen, M. Grote, and M. Carrington (1995). Discordant patterns of linkage disequilibrium of the peptide transporter loci within the HLA class II region. *Am J Hum Genet 57*, 1436–1444.

Knapp, M. (1999a). A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet 64*, 1177–1185.

Knapp, M. (1999b). The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet 64*, 861–870.

Knapp, M. (1999c). Using exact p values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib tranmission/disequilibrium test. *Am J Hum Genet 65*, 1208–1210.

Knapp, M., G. Wassmer, and M. P. Baur (1995). The relative efficiency of the hardy-weinberg equilibrium-likelihood and the conditional on parental genotype-likelihood methods for candidate-gene association studies. *Am J Hum Genet 57*, 1476–1485.

Knight, K. (2000). *Mathematical Statistics*. Chapman & Hall /CRC.

Koch, H., J. McClay, E.-W. Loh, S. Higuchi, J.-H. Zhao, P. Sham, D. Ball, and I. W. Craig (2000). Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. *Hum Mol Genet 9*, 2993–2999.

Koeleman, B. P. C., F. Dudbridge, H. Cordell, and J. A. Todd (2000). Adaption of the extended tranmission/disequilibrium test to distinguish disease association of multiple loci: the conditional extended transmission/disequilibrium test. *Ann Hum Genet 64*, 207–213.

Kong, A., N. Cox, M. Frigge, and M. Irwin (1993). Sequential imputation and multipoint linkage analysis. *Genet Epidemiol 10*, 483–488.

Kong, A. and N. J. Cox (1997). Allele-sharing models: Lod scores and accurate linkage tests. *Am J Hum Genet 61*, 1179–1188.

Konigsberg, L. W., C. M. Kammerer, and J. W. MacCluer (1989). Segregation analysis of quantitative traits in nuclear families: comparison of three program packages. *Genet Epidemiol 6*, 713–726.

Krone, S. M. and C. Neuhauser (1997). Ancestral processes with selection. *Theor Popul Biol 51*, 210–237.

Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nat Genet 17*(1), 21–4.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of complex disease genes. *Nat Genet 22*, 139–145.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly, and E. S. Lander (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet 58*, 1347–1363.

Kruglyak, L. and E. S. Lander (1995). High-resolution genetic mapping of complex traits. *Am J Hum Genet 56*, 1212–1223.

Kruglyak, L. and E. S. Lander (1998). Faster multipoint linkage analysis using Fourier transforms. *J Comp Biol 5*, 1–7.

Kuhner, M. K., J. Yamato, and J. Felsenstein (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics 140*, 1421–1430.

Kuhner, M. K., J. Yamato, and J. Felsenstein (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics 149*, 429–434.

Kwok, P.-Y. and Z. Gu (1999). Single nucleotide polymorphism libraries: why and how are we building them ? *Molecular Medicine Today 5*, 538–543.

Laan, M. and S. Pääbo (1997). Demographic history and linkage disequilibrium in human populations. *Nat Genet 17*, 435–373.

Laird, N. M., S. Horvath, and X. Xu (2000). Implementing a unified approach to family-based tests of association. *Genet Epidemiol 19 (supplement)*, 36–42.

Lake, S. L., D. Blacker, and N. M. Laird (2000). Family-based tests of association in the presence of linkage. *Am J Hum Genet 67*, 1515–1525.

Lalouel, J. M. and N. E. Morton (1981). Complex segregation analysis with pointers. *Hum Hered 31*, 312–321.

Lalouel, J. M., D. C. Rao, N. E. Morton, and R. C. Elston (1983). A unified model for complex segregation analysis. *Am J Hum Genet 35*, 816–826.

Lalouel, J. M. and S. Yee (1980). POINTER: Computer programs for complex segregation

analysis of nuclear families with pointers. population genetics laboratory. Technical Report 3, University of Hawaii.

Lander, E. and L. Kruglyak (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet 11*, 241–247.

Lander, E. and N. Schork (1994). Genetic dissection of complex traits. *Science 265*, 2037–48.

Lander, E. S. (1995). Mapping heredity: using probabilistic models and algorithms to map genes and genomes. In E. S. Lander and M. S. Waterman (Eds.), *Calculating the Secrets of Life*, pp. 25–55. Washington: National Academy Press.

Lander, E. S. and D. Botstein (1987a). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science 236*, 1567–1570.

Lander, E. S. and D. Botstein (1987b). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science 236*, 1567–1570.

Lander, E. S. and P. Green (1987). Construction of multilocus genetic maps in humans. *Proceedings of the National Academy of Sciences USA 84*, 2363–2367.

Lander, E. S. and M. S. Waterman (1995). *Calculating the Secretes of Life*. Washington: National Academy Press.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Statistics for Biology and Health. New York: Springer-Verlage.

Lange, K., M. Boehnke, and D. Weeks (1987). Programs for pedigree analysis: Mendel, fisher, and dgene. *Genet. Epidemiol. 5*, 473–476.

Lange, K. and R. C. Elston (1975). Extensions to pedigree analysis. I. likelihood calculations for simple and complex pedigrees. *Hum Hered 25*, 95–105.

Lange, K. and E. Sobel (1991). A random walk method for computing genetic location scores. *Am J Hum Genet 49*, 1320–1334.

Lange, K., D. E. Weeks, and M. Boehnke (1988). Programs for pedigree analysis: MENDEL, FISHER, dGENE. *Genet Epidemiol 5*, 471–72.

Lathrop, G. M., J. M. Lalouel, C. Julier, and J. Ott (1984). Strategies for multilocus linkage analysis in humans. In *Proceedings of the National Academy of Science USA*, Volume 81, pp. 3443–3446.

Lazzeroni, L. C. (1998). Linkage disequilibrium and gene mapping: an empirical least-squares approach. *Am J Hum Genet 62*, 159–170.

Lazzeroni, L. C. (2001). A chronology of fine-scale gene mapping by linkage disequilibrium. *Stat Meth Med Res 10*, 57–76.

Lazzeroni, L. C. and K. Lange (1997). Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. *Ann Statist 25*, 138–168.

Lazzeroni, L. C. and K. Lange (1998). A conditional inference framework for extending the tranmission/disequilibrium test. *Hum Hered 48*, 67–81.

Lee, H. and D. O. Stram (1996). Segregation analysis of continuous phenotypes by using higher sample moments. *Am J Hum Genet 58*, 213–224.

Lee, H., D. O. Stram, and D. C. Thomas (1993). A generalized estimating equation approach to fitting major gene models in segregation analysis of continuous phenotypes. *Genet Epidemiol 10*, 61–74.

Lee, J. K. and D. C. Thomas (2000). Performance of Markov chain Monte Carlo approaches for mapping genes in oligogenic models with an unknown number of loci. *Am J Hum Genet 67*, 1232–1250.

Lehesjoki, A. E., M. Koskiniemi, R. Norio, S. Tirrito, P. Sistonen, E. Lander, and A. de la Chapelle (1993). Localization of the epm1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. *Hum Mol Genet 2*, 1229–1234.

Lewis, P. O. and D. Zaykin (1999). Genetic data analysis: Computer program for the analysis of allelic data, http://chee.umn.edu/gda.

Liang, K.-Y., Y.-F. Chiu, and T. Beaty (2001). A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. *Hum Hered 51*, 64–78.

Liang, K.-Y. and P. J. Rathouz (1999). Hypothesis testing under mixture models: application to genetic linkage analysis. *Biometrics 55*, 65–74.

Lin, M., J. H. Zhao, D. Curtis, and P. C. Sham (1997). Power comparisons of nonparametric linkage tests. *Am J Med Genet (NeuroPsychiatric Genetics) 74*, 601.

Lin, S. (1996). Monte Carlo methods in genetic analysis. In T. Speed and M. S. Waterman (Eds.), *Genetic Mapping and DNA Sequencing*, The IMA Volumes in Mathematics and its Applications, pp. 15–38. Springer.

Lin, S., E. A. Thompson, and E. Wijsman (1993). Achieving irreducibility of the Markov Chain Monte Carlo method applied to pedigree data. *IMA Journal of Mathematics Applied in Medicine and Biology 10*, 1–17.

Lindblad-Toh, K., E. Winchester, M. J. Daly, D. G. Wang, J. N. Hirschhorn, J.-P. Laviolette, K. Ardlie, D. E. Reich, E. Robinson, P. Sklar, N. Shah, D. Thomas, J.-B. Fan, T. Gingeras, J. Warrington, N. Patil, J. H. Thomas, and E. S. Lander (2000). Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet 24*, 381–386.

Liu, J. S., C. Sabatti, J. Teng, B. J. Keats, and N. Risch (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res 11*, 1716–1724.

Long, A. D. and C. H. Langley (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res 9*, 720–731.

Long, J. C., R. C. Williams, and M. Unbanek (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet 56*, 799–810.

Longmate, J. A. (2001). Complexity and power in case-control association studies. *Am J Hum Genet 68*, 1229–1237.

MacCluer, J. W. (1967). Monte Carlo methods in human population genetics: a computer model incorporating age-specific birth and dath rates. *Am J Hum Genet 19*, 303–312.

MacLean, C. J., R. B. Martin, P. C. Sham, H. Wang, R. E. Straub, and K. S. Kendler (2000). The trimmed-haplotype test for linkage disequilibrium. *Am J Hum Genet 66*, 1062–1075.

MacLean, C. J., N. E. Morton, R. C. Elston, and S. Yee (1976). Skewness in commingled distributions. *Biometrics 32*, 695–699.

Martin, E. R., N. L. Kaplan, and B. S. Weir (1997). Tests for linkage and association in nuclear families. *Am J Hum Genet 61*, 439–448.

Martin, E. R., S. A. Monks, L. L. Warren, and N. L. Kaplan (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet 67*, 146–154.

McCarthy, M. I., L. Kruglyak, and E. S. Lander (1998). Sib-pair collection strategies for complex diseases. *Genet Epidemiol 15*, 317–340.

McGinnis, R. E. (1998). Hidden linkage: a comparison of the affected sib pair (asp) test and transmission/disequilibrium test (TDT). *Ann Hum Genet 62*, 159–179.

McGinnis, R. E. (2000). General equations for p_t, p_s and the power of the TDT and the affected-sib-pair test. *Am J Hum Genet 67*, 1340–1347.

Mckeigue, P. M. (1997). Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am J Hum Genet 60*, 188–196.

McKeigue, P. M. (2000). Efficiency of estimation of haplotype frequencies: use of marker phenotypes of unrelated individuals versus counting of phase-known gametes. *Am J Hum Genet 67*, 1626–1627.

McPeek, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol 16*(3), 225–249.

McPeek, M. S. and A. Strahs (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet 65*, 858–875.

Monks, A. A. and N. L. Kaplan (2000). Removing the sampling restrictions from family-basd tests of association for a quantitative trait locus. *Am J Hum Genet 66*, 576–592.

Monks, S., N. Kaplan, and B. Weir (1998). A comparative study of sibship tests of linkage and/or association. *Am J Hum Genet 63*, 1507–1516.

Moore, K. J. and D. L. Nagle (2000). Ccomplex trait analysis in the mouse: the strengths, the limitations and the promise yet to come. *Ann Rev Genet 34*, 653–686.

Morris, A. P., R. N. Curnow, and J. C. Whittaker (1997). Randomization tests of disease-marker asssociations. *Ann Hum Genet 61*, 49–60.

Morris, A. P., J. C. Whittaker, and D. J. Balding (2000). Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am J Hum Genet 67*, 155–169.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet 7*, 277–318.

Morton, N. E. (1982). Trials of segregation analysis by deterministic and macro simulation. *Am J Hum Genet 34*, 187A.

Morton, N. E. (2000). Unsolved problems in genet epidemiol. *Hum Hered 50*, 5–13.

Morton, N. E. and A. Collins (1998). Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci 95*, 11389–11393.

Morton, N. E. and C. J. MacLean (1974). Analysis of family resemblance. III. Complex segregation of quantitative traits. *Am J Hum Genet 26*, 489–503.

Morton, N. E., D. C. Rao, and J. M. Lalouel (1983). *Methods in Genetic Epidemiolgy*. Karger.

Mukhopadhyay, N., L. Almasy, M. Schroeder, W. P. Mulvihill, and D. E. Weeks (1999). Mega2, a data-handling program for facilitating genetic linkage and association analyses. *Am J Hum Genet 65*, A436.

Müller-Myhsok, B. and L. Abel (1997). Genetic analysis of complex traits. *Science 275*, 1329.

Murray, R. (1997). Schizophrenia. In R. Murray, P. Hill, and P. McGuffin (Eds.), *The Essentials of Postgraduate Psychiatry* (Third ed.)., pp. pp281–309. Cambridge University Press.

Nettleton, D. and R. W. Doerge (2000). Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics 56*, 52–58.

Neuhauser, C. and S. M. Krone (1997). The genealogy of samples in models with selection. *Genetics 145*, 519–534.

Nevanlinna, H. R. (1980). Genetic markers in Finland. *Haematologia 13*, 65–74.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson, J. Stengard, V. Salomaa, E. Vartiainen, E. Boerwinkle, and C. F. Sing (1998). DNA sequence in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet 19*, 233–240.

Nielsen, D. M., M. Ehm, and B. Weir (1998). Detecting marker-disease association by testing for hardy-weinberg disequilibrium at a marker locus. *Am J Hum Genet 63*, 1531–1540.

Nielsen, D. M. and B. S. Weir (1999). A classical setting for associations between markers and loci affecting quantitative traits. *Genetical Research 74*, 271–277.

Nielsen, D. M. and B. S. Weir (2001). Association studies under general disease models. *Theor Popul Biol 60*, 253–263.

Nielsen, R. (1997). A likelihood approach to population samples of microsatellite alleles. *Genetics 146*, 711–716.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics 154*, 931–942.

Nyholt, D. R. (2000). All LODs are not create equal. *Am J Hum Genet 67*, 282–288.

O'Connell, J. R. (2001). Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. *Hum Hered 51*, 226–240.

O'Connell, J. R. and D. E. Weeks (1995). The vitesse algorithm for rapid exact multilocus linkage analysis via genotype set-recording and fuzzy inheritance. *Nat Genet 11*, 402–408.

Ohadi, M., M. R. A. Lalloz, P. C. Sham, J. H. Zhao, A. M. Dearlove, C. Shiach, S. Kinsey, M. Rhodes, and M. D. Layton (1999). Localization of a gene for familial hemophagocytic lymphohistocystosis at chromosome 9q21.3-22 by homozygosity mapping. *Am J Hum Genet 64*, 165–171.

Ohta, T. and M. Kimura (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res 22*, 210–204.

Olson, J. M., J. S. Witte, and R. C. Elston (1999). Genetic mapping of complex traits. *Stat Med 18*, 2961–2981.

Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. *Am J Hum Genet 26*, 588–597.

Ott, J. (1977). Couning methods (em algorithm) in human pedigree analysis: linkage and segregation analysis. *Ann Hum Genet 40*, 443–454.

Ott, J. (1985). *Analysis of Human Genetic Linkage* (First ed.). Johns Hopkins University Press.

Ott, J. (1989a). Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci USA 86*, 4175–4178.

Ott, J. (1989b). Statistical properties of the haplotype relative risk. *Genet Epidemiol 6*, 127–130.

Ott, J. (1991). *Analysis of Human Genetic Linkage* (Revised ed.). Johns Hopkins University Press.

Ott, J. (1992). Strategies for characterizing highly polymorphic markets in human gene mapping. *Am J Hum Genet 51*, 283–290.

Ott, J. (1999). *Analysis of Human Genetic Linkage* (Third ed.). Baltimore and London: The Johns Hopkins University Press.

Ott, J. and D. Rabinowitz (1997). The effect of marker heterozygosity on the power to detect linkage disequilibrium. *Genetics 147*(2), 927–30.

Page, G. P., C. I. Amos, and E. Boerwinkle (1998). The quantitative lod score: test statistic and sample size for exclusion and linkage of quantitative traits in human sibships. *Am J Hum Genet 62*, 962–968.

Patnaik, P. B. (1949). The non-central χ^2 and f-distributions and their applications. *Biometrika 36*, 202–232.

Pawlowitziki, I.-H., J. H. Edwards, and E. A. Thompson (1997). *Genetic Mapping of Disease Genes*. London: Academic Press.

Pennisi, E. (1998). A closer look at SNPs suggests difficulties. *Science 281*, 1787–1789.

Penrose, L. S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics 6*, 133–138.

Pietro, L. and N. Goldman (1998). Models of molecular evolution and phylogeny. *Genome Res 8*, 1233–1244.

Ploughman, L. M. and M. Boehnke (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *Am J Hum Genet 44*, 543–551.

Pratt, S. C., M. J. Daly, and L. Kruglyak (2000). Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *Am J Hum Genet 66*, 1153–1157.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes: The Art of Scientific Computing* (2nd ed.). Cambridge: Cambridge University Press.

Pritchard, J. K. and M. Przeworski (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet 69*, 1–14.

Pritchard, J. K. and N. A. Rosenberg (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet 65*, 220–228.

Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly (2000). Association mapping in structural populations. *Am J Hum Genet 67*, 170–181.

Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum Hered 47*, 342–350.

Rabinowitz, D. and N. Laird (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structures and arbitrary missing marker information. *Hum Hered 50*, 211–223.

Rannala, B. and M. Slatkin (1998). Likelihood analysis of disequilibrium mapping, and related problems. *Am J Hum Genet 62*, 459–473.

Rao, D. C. and R. C. Elandt-Johnson (1971). Complex segregation analysis. *Am J Hum Genet 23*, 325–326.

Reich, D. E. and D. B. Goldstein (2001). Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol 20*, 4–16.

Risch, N. (1984). Segregation analysis incorporating linkage markers. I. single-locus models with an application to type i diabetes. *Am J Hum Genet 36*, 363–386.

Risch, N. (1989). Linkage detection tests under heterogeneity. *Genet Epidemiol 6*, 473–480.

Risch, N. (1990a). Linkage strategies for genetically complex traits: I. multilocus models. *Am J Hum Genet 46*, 222–8.

Risch, N. (1990b). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet 46*, 242–253.

Risch, N. (1990c). Linkage strategies for genetically complex traits. II.The power of affected relative pairs. *Am J Hum Genet 46*, 229–241.

Risch, N. (1997). Evolving methods in genet epidemiol. II. genetic linkage from an epidemiologic perspective. *Epidemiologic Reviews 19*, 24–32.

Risch, N. and L. Giuffra (1992). Model misspecification and multipoint linkage analysis. *Hum Hered 42*, 77–92.

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

Risch, N. and J. Teng (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. *Genome Res 8*, 1273–1288.

Risch, N. and H. Zhang (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science 268*, 1584–1589.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature 405*, 847–856.

Rubinstein, P., M. Walker, C. Carpenter, C. Carrier, J. Krassner, C. Falk, and F. Ginsberg (1981). Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups. *Human Immunology 3*, 384.

S.A.G.E (1998). Statistical Analysis for Genetic Epidemiology, Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University.

Sandkuijl, L. A. (1989). Analysis of affected sib-pairs using information from extended families. In *Multipoint mapping and linkage based upon affected pedigree members: Genetic Analysis Workshop 6*. New York: Alan R Liss.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics 53*, 1253–1261.

Schaid, D. J. (1996). General score tests for association of genetic markers with disease using cases and their parents. *Genet Epidemiol 13*, 423–449.

Schaid, D. J. (1998). Statistical genetics '98: Transmission disequilibrium, family controls, and great expectations. *Am J Hum Genet 63*, 935–941.

Schaid, D. J. (1999a). Case-parents design for gene-environment interaction. *Genet Epidemiol 16*(3), 261–273.

Schaid, D. J. (1999b). Likelihoods and tdt for the case-parents design. *Genet Epidemiol 16*(3), 250–260.

Schaid, D. J. and C. Rowland (1998). Use of parents, sibs, and unrelated controls for detection of association between genetic markers and disease. *Am J Hum Genet 63*, 1492–1506.

Schaid, D. J. and C. M. Rowland (2000). Robust transmission regression models for linkage and association. *Genet Epidemiol 19*, 78–84.

Schaid, D. J. and S. S. Sommer (1993). Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet 53*, 1114–1126.

Schaid, D. J. and S. S. Sommer (1994). Comparison of statistics for candidate-gene association studies using cases and parents. *Am J Hum Genet 55*, 402–409.

Schneider, S., J. M. Kueffer, D. Roessli, and L. Excoffier (1998). Arlequin: a software for population genetic analysis, http://anthopologie.unige.chh/arlequin.

Schork, N. J. (1992). Detection of genetic heterogeneity for complex quantitative phenotypes. *Genet. Epidemiol. 9*, 207–223.

Schork, N. J. (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Am J Hum Genet 55*, 1306–1319.

Schork, N. J., L. R. Cardon, and X. Xu (1998). The future of genet epidemiol. *Trends in Genetics 14*(7), 266–272.

Schull, W. J. (1967). On the use of computers in human genetics: an introduction. *Am J Hum Genet 19*, 183–188.

Scott, W. K., M. A. Pericak-Vance, and J. L. Haines (1997). Genetic analysis of complex diseases. *Science 275*, 1327.

Searls, D. B. (2000). Bioinformatics tools for whole genomes. *Annual Review of Genomics and Human Genetics 1*, 251–279.

Self, S. G. and K. Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Amer Stat Assoc 82*, 605–610.

Seltman, H., K. Roeder, and B. Devlin (2001). Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet 68*, 1250–1263.

Sengul, H., D. E. Weeks, and E. Feingold (2001). A survey of affected-sibship statistics for nonparametric linkage analysis. *Am J Hum Genet 69*, 179–190.

Sham, P. C. (1997). Transmission/disequilibrium tests for multiallelic loci. *Am J Hum Genet 61*, 774–778.

Sham, P. C. (1998). *Statistics in Human Genetics*. Arnold Applications of Statistics Series. London: Edward Arnold.

Sham, P. C., S. S. Cherny, S. Purcell, and J. K. Hewitt (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet 66*, 1616–1630.

Sham, P. C. and D. Curtis (1995a). An extended tranmission/disequilibrium test (TDT) for multi-allelic marker loci. *Ann Hum Genet 59*, 323–336.

Sham, P. C. and D. Curtis (1995b). Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann Hum Genet 59*, 97–105.

Sham, P. C., M.-W. Lin, J. H. Zhao, and D. Curtis (2000). Power comparison of parametric and nonparametric linkage tests in small pedigrees. *Am J Hum Genet 66*, 1661–1668.

Sham, P. C. and S. Purcell (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am J Hum Genet 68*, 1527–1532.

Sham, P. C. and J. H. Zhao (1998). Linkage analysis using affected sib-pairs. In M. J. Bishop (Ed.), *Guide to Human Genome Computing* (Second ed.)., pp. 75–88. Academic Press.

Sham, P. C., J. H. Zhao, and D. Curtis (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann Hum Genet 61*, 61–69.

Sham, P. C., J. H. Zhao, and D. Curtis (2000). The effect of marker characteristics on the power to detect linkage disequilibrium due to single or multiple ancestral mutations. *Ann Hum Genet 64*, 161–169.

Sheehan, N. and A. Thomas (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics 49*, 163–175.

Shifman, S. and A. Darvasi (2001). The value of isolated populations. *Nat Genet 28*, 309–310.

Shih, M.-C. and A. S. Whittemore (2001). Allele-sharing among affected relatives: non-parametric methods for identifying genes. *Stat Meth Med Res 10*, 27–55.

Slager, S., J. Huang, and V. Vieland (2001). Power comparisons between the TDT and two likelihood-based methods. *Genet Epidemiol 20*, 192–209.

Slatkin, M. (1999). A coalescent view of population structure. In R. S. Singh and C. B. Krimbas (Eds.), *Evolutionary Genetics: From Molecules to Morphology*, Volume I, pp. 418–429. Cambridge University Press.

Smith, C. A. B. (1953). Detection of linkage in human genetics. *J Roy Stat Soc B 15*, 153–192.

Smith, C. A. B. (1963). Testing for heterogeneity of recombination fraction values in human genetics. *Ann Hum Genet 27*, 175–182.

Sobel, E. and K. Lange (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet 58*, 1323–1337.

Speed, T. and M. S. Waterman (1996). *Genetic Mapping and DNA Sequencing*, Volume 81 of *The IMA Volumes in Mathmatics and its Applications*. Springer.

Speed, T. P. (1996). What is a genetic map function ? In T. P. Speed and M. S. Waterman (Eds.), *Genetic Mapping and DNA Sequencing*, Volume 81 of *The IMA Volumes in Mathematics and its Applications*, pp. 65–88. Springer.

Speer, M., J. D. Terwilliger, and J. Ott (1992). A chromosome-based method for rapid computer simulation. *Am J Hum Genet 51*, A202.

Spielman, R. S. and W. J. Ewens (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet 62*, 450–458.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet 52*, 506–516.

Strauch, K., R. Fimmers, T. Kurz, K. A. Deichmann, T. F. Wienker, and M. P. Baur (2000). Parametric and nonparametric multipoint linkage analysis with imprinting ant two-locus-trait models: applications to mite sensitization. *Am J Hum Genet 66*, 1945–1957.

Suarez, B. K., J. P. Rice, and T. Reich (1978). The generalised sib-pair IBD distribution: Its use in the detection of linkage. *Ann Hum Genet 42*, 87–94.

Suarez, B. K. and P. van Eerdewegh (1984). A comparison of three affected-sib-pair scoring methods to detect hla-linked disease susceptibility genes. *Am J Med Genet 18*, 135–146.

Subrahmanyan, L., M. A. Eberle, A. G. Clark, L. Kruglyak, and D. A. Nickerson (2001). Sequence variation and linkage disequilibrium in the human T-cell receptor (TCRB) locus. *Am J Hum Genet 69*, 381–395.

Sun, F., W. D. Flanders, Q. Yang, and M. J. Khoury (1999). Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *American Journal of Epidemiology 150*, 97–104.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol 2*, 125–141.

Tavaré, S., D. J. Balding, R. Griffiths, and P. Donnelly (1997). Inferring coalescence times from DNA sequence data. *Genetics 145*, 505–518.

Teng, J. and N. Risch (1999). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. individual genotyping. *Genome Res 9*, 234–241.

Teng, J. and D. Siegmund (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics 54*, 1247–1265.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet 56*, 777–787.

Terwilliger, J. D., D. Y. Ding, and J. Ott (1992). On the relative importance of marker heterozygosity and intermarker distance in gene mapping. *Genomics 13*, 951–956.

Terwilliger, J. D. and H. H. H. Goöring (2000). Gene mapping in the 21st centuries: statistical methods, data analysis, and experimental design. *Human Biology 72*, 63–132.

Terwilliger, J. D. and J. Ott (1992). A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered 42*, 337–346.

Terwilliger, J. D. and J. Ott (1994). *Handbook of Human Genetic Linkage*. Baltimore: The Johns Hopkins University Press.

Terwilliger, J. D., M. Speer, and J. Ott (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet. Epidemiol. 10*, 217–224.

Thomas, D. C. and V. Cortessis (1992). A Gibbs sampling approach to linkage analysis. *Hum Hered 42*, 63–76.

Thompson, E. A. (1995). Monte Carlo in genetic analysis. Technical Report 294, University of Washington.

Thompson, E. A. (1996). Statistical genetics. In P. Armitage and H. David (Eds.), *Advances in Biometry: The last 50 years*, pp. 263–285. New York: Wiley.

Thompson, E. A. (2000a). MCMC estimation of multi-locus genome sharing and multipoint gene location scores. *International Statistical Review 68*, 53–73.

Thompson, E. A. (2000b). Monte Carlo methods on genetic structures. In O. E. Barndorff-Nielson, D. R. Cox, and C. Klüppelberg (Eds.), *Complex Stochastic Systems*, Séminaire Européan de Statistique, pp. 179–222. London: Chapman and Hall.

Thompson, E. A. (2000c). *Statistical Inferences from Genetic Data on Pedigrees*, Volume 6 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Beachwood, Ohio: IMS.

Thompson, E. A. (2001). PANGAEA: pedigree analysis for genetics, http://www.stat.washington.edu/thompson/genepi/pangaea.shtml.

Thompson, E. A. and S. W. Guo (1991). Evaluation of likelihood ratios for complex genetic models. *IMA Journal of Mathematics Applied in Medicine and Biology 8*,

149–169.

Thompson, E. A., S. Lin, A. Olshen, and E. Wijsman (1993). Monte Carlo analysis of a large hypercholesterolemia pedigree. *Genet Epidemiol 10*, 677–682.

Thompson, E. A. and J. V. Neel (1997). Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am J Hum Genet 60*, 197–204.

Thomson, G. (1988). HLA disease associations: models for insulin dependent diabetes mellitus and the study of complex human genetic disorders. *Ann Rev Genet 22*, 31–50.

Thomson, G. (1995a). Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. *Am J Hum Genet 57*, 474–486.

Thomson, G. (1995b). Mapping disease genes: family-based association studies. *Am J Hum Genet 57*, 487–498.

Thomson, G. and M. S. Esposito (1999). The genetics of complex diseases. *Trends in Cell Biology 9*, M17–20.

Toivonen, H. T. T., P. Onkamo, K. Vakso, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere (2000). Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet 67*, 133–145.

Trégouët, D. A., C. Pallaud, C. Sass, S. Visvikis, and L. Tiret (2001). Sample size calculations for classical association and TDT-type methods using family data. *Ann Hum Genet 65*, 293–312.

Trégouët, D. A. and L. Tiret (2000). Applications of the estimating equations theory to genetic epidemiology: a review. *Ann Hum Genet 64*, 1–14.

Umbach, D. M. and C. R. Weinberg (2000). The use of case-parent to study joint effects of genotype and exposure. *Am J Hum Genet 66*, 251–261.

Van den Oord, E. (2000). Testing for linkage disequilibrium, maternal effects, and imprinting with (in)complete case-parent triads, by use of the computer program LEM. *Am J Hum Genet 66*, 335–338.

van den Oord, E. J. C. G. (1999). A comparison between different designs and tests to detect QTLs in association studies. *Behav Genet 29*, 245–256.

Waldman, I. D., B. F. Robinson, and D. C. Rowe (1999). A logistic regression based

extension of the tdt for continuous and categorical traits. *Ann Hum Genet 63*, 329–340.

Waller, L. A., B. W. Turnbull, and H. J. Michael (1995). Obtaining distribution functions by numerical inversion of characteristic functions with applications. *The American Statistician 49*(4), 346–350.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science 280*, 1077–1082.

Weeks, D. E. and K. Lange (1988). The affected-pedigree-member method of linkage analysis. *Am J Hum Genet 42*, 315–326.

Weeks, D. E. and G. M. Lathrop (1995). Polygenic disease: methods for mapping complex disease traits. *Trends in Genetics 11*, 513–519.

Weeks, D. E., J. Ott, and G. M. Lathrop (1990). SLINK: a general simulation program for linkage analysis (abstract). *Am J Hum Genet 47*, A204.

Weinberg, C. R. (1999a). Analysis of genetic studies with partial triads. *Am J Hum Genet 64*, 1186–1193.

Weinberg, C. R. (1999b). Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet 65*, 229–235.

Weinberg, C. R., A. J. Wilcox, and R. T. Lie (1998). A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that mat be subjected to parental imprinting. *Am J Hum Genet 62*, 969–978.

Weir, B. S. (1990). *Genetic Data Analysis*. Sunderland, Massachusetts: Sinauer Associates, Inc.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, Massachusetts: Sinauer Associates, Inc.

Weir, B. S. and C. C. Cockerham (1978). Testing hypothesis about linkage disequilibrium with multiple alleles. *Genetics 88*, 633–642.

Weir, B. S. and C. C. Cockerham (1979). Estimation of linkage disequilibrium in randomly mating populations. *Heredity 42*, 105–111.

Weiss, K. M. (1995). *Genetic Variation and Human Disease: Principles and Evolutionary Approaches*. Cambridge Studies in Biological Anthropology. Cambridge University Press.

Whelan, S., P. Liò, and N. Goldman (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics 17*, 262–272.

Whittaker, J. C. and C. M. Lewis (1999). Power comparisons of the transmission/disequilibrium test and sib-transmission/disequilibrium-test statistics. *Am J Hum Genet 65*, 578–580.

Whittaker, J. C. and D. J. Thompson (1999). Finite-sample properties of family-based association tests. *Am J Hum Genet 64*, 910–915.

Whittemore, A. S. (1996). Genome scanning for linkage: an overview. *Am J Hum Genet 59*, 704–716.

Whittemore, A. S. and J. Halpern (1994a). A class of tests for linkage using affected pedigree members. *Biometrics 50*, 118–127.

Whittemore, A. S. and J. Halpern (1994b). Probability of gene identity by descent: computation and applications. *Biometrics 50*, 109–117.

Whittemore, A. S. and I.-P. Tu (2000). Detection of disease genes by using family data: I. likelihood-based theory. *Am J Hum Genet 66*, 1328–1340.

Wijsman, E. M. and C. I. Amos (1997). Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. *Genet Epidemiol 14*, 719–735.

Williams, J. T. and J. Blangero (1999). Power of variance component linkage analysis to detect quantitative trait loci. *Ann Hum Genet 63*, 545–563.

Williams, J. T., P. V. Eerdewegh, L. Almasy, and J. Blangero (1999). Joint multipoint linkage analysis of multivariate quantitative and quantitative traits: I. likelihood formulation and simulation results. *Am J Hum Genet 65*, 1134–1147.

Wilson, A. F., J. E. Bailey-Wilson, E. W. Pugh, and S. A. J. M. (1996). Genometric analysis simulation program. *Am J Hum Genet 59*, A193.

Wilson, I. J. and D. J. Balding (1998). Genealogical inference from microsatellite data. *Genetics 150*, 499–510.

Wiuf, C. (2000). On the genealogy of a sample of neutral rare alleles. *Theor Popul Biol 58*, 61–75.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann Hum Genet 19*, 251–253.

Workman, P. L. and J. D. Niswander (1970). Population studies on southwestern indian tribes. II. local genetic differentiation in the papago. *Am J Hum Genet 22*, 24–39.

Wright, A. F., A. D. Carothers, and M. Pirastu (1999). Population choice in mapping genes for complex diseases. *Nat Genet 23*, 397–404.

Wright, F. A. (1997). The phenotypic difference discards sib-pair QTL linkage information. *Am J Hum Genet 60*, 740–742.

Xie, X. and J. Ott (1993). Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet 53*, 1107.

Xiong, M. and S.-W. Guo (1997). Fine-scale genetic mapping based on linkage disequilibrium. *Am J Hum Genet 60*, 1513–1531.

Xiong, M. and L. Jin (2000). Combined linkage and linkage disequilibrium mapping for genome screens. *Genet Epidemiol 19*, 211–234.

Zapata, C., G. Alvarez, and C. Carollo (1997). Approximate variance of the standardized measure of gametic disequilibrium D'. *Am J Hum Genet 61*, 771–774.

Zapata, C., C. Carllos, and S. Rodríguez (2001). Sampling variance and distribution of the the D' measure of overall gametic disequilibrium between multiallelic loci. *Ann Hum Genet 65*, 395–406.

Zaykin, D., L. Zhivotovsky, and B. S. Weir (1995). Exact tests for association between alleles at arbitrary numbers of loci. *Genetica 96*, 169–178.

Zhao, H. (2000). Family-based association studies. *Stat Meth Med Res 9*, 563–587.

Zhao, H., M. S. McPeek, and T. P. Speed (1995a). A statistical analysis of chromatid interference. *Genetics 139*, 1057–1065.

Zhao, H., M. S. McPeek, and T. P. Speed (1995b). A statistical analysis of crossover interference using the chi-square model. *Genteics 139*, 1045–1046.

Zhao, H., A. J. Pakstis, J. R. Kidd, and K. K. Kidd (1999). Assessing linkage disequilibrium in a complex genetic system. I. overall deviation from random association. *Ann Hum Genet 63*, 167–179.

Zhao, H., S. Zhang, K. R. Merikangas, M. Trixler, D. B. Wildenauer, F. Sun, and K. K. Kidd (2000). Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet 67*, 936–946.

Zhao, J. H., D. Curtis, and P. C. Sham (2000). Model-free analysis and permutation tests for allelic associations. *Hum Hered 52*, 133–139.

Zhao, J. H. and P. C. Sham (1998). A method for calculating probability convolution using ternary numbers with application in the determination of twin zygosity. *Computional Statistics and Data Analysis 28*, 225–232.

Zhao, J. H. and P. C. Sham (2002). Faster haplotype frequency estimation using unrelated subjects. *Hum Hered 53*, 36–41.

Zhao, J. H., P. C. Sham, and D. Curtis (1999). A program for the Monte Carlo evaluation of significance of the extended transmission/disequilibrium test. *Am J Hum Genet 64*, 1484–1485.

Zhao, L. P. (1994). Segregation analysis of human pedigrees using estimating equations. *Biometrika 81* (4), 197–209.

Zhao, L. P., C. Aragaki, L. Hsu, and F. Quiaoit (1998). Mapping of complex traits by single-nucleotide polymorphism. *Am J Hum Genet 63*, 225–240.

Zhao, L. P., L. Hsu, O. Davidov, J. Potter, R. C. Elston, and R. L. Prentice (1997). Population-based family study designs: An interdisciplinary research framework for genetic epidemiology. *Genet Epidemiol 14*, 365–388.

Zhu, X. and R. C. Elston (2001). Transmission/disequilibrium tests for quantitative traits. *Genet Epidemiol 20*, 57–74.

Zhu, X., R. C. Elston, and R. S. Cooper (2001). Testing quantitative traits for association and linkage in the presence or absence of parental data. *Hum Hered 51*, 183–191.

Ziegler, A. and J. Hebebrand (1998). Sample size calculations for linkage analysis using extreme sib pairs based on segregation analysis with the quantitative phenotype body weight as an example. *Genet Epidemiol 15*(6), 577–593.

Zöllner, S. and A. von Haeseler (2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am J Hum Genet 66*, 615–628.

Common journal abbreviations

Abbreviation	Full name
Am J Hum Genet	American Journal of Human Genetics
Am J Med Genet	American Journal of Medical Genetics
Ann Hum Genet	Annals of Human Genetics
Ann Math Stat	Annals of Mathematical Statistics
Ann Rev Genet	Annual Review of Genetics
Ann Statist	Annals of Statistics
Behav Genet	Behavior Genetics
Epidemiol Rev	Epidemiological Reviews
Genet Epidemiol	Genetic Epidemiology
Genet Res	Genetic Research
Genome Res	Genome Research
Hum Biol	Human Biology
Hum Hered	Human Heredity
Hum Mol Genet	Human Molecular Genetics
Int Stat Rev	International Statistical Review
J Amer Stat Assoc	Journal of American Statistical Association
J Comp Biol	Journal of Computational Biology
J Hered	Journal of Heredity
J Roy Stat Soc	Journal of Royal Statistical Society
J Stat Plan Infer	Journal of Statistical Planning and Inference
Nat Genet	Nature Genetics
Proc Natl Acad Sci	Proceedings of National Academy of Science
Stat Med	Statistics in Medicine
Stat Meth Med Res	Statistical Methods in Medical Research
Theor Popul Biol	Theoretical Population Biology