Abstract

LI, LI. Disease Gene Mapping in General Pedigrees (under the direction of Dr. Bruce S. Weir and Dr. Sharon R. Browning)

Disease gene mapping is one of the main focuses of genetic epidemiology and statistical genetics. This dissertation explores some methods and algorithms in this area, especially in pedigrees. The first chapter gives an introduction to human genetics and disease gene mapping. Existing linkage and association methods are introduced and compared. Probabilities of genotypic data from multiple linked marker loci on related individuals are used as likelihoods of gene locations for gene-mapping, or as likelihoods of other parameters of interest in human genetics. With the recent development in genetics and molecular biology techniques, large-scale marker data has become available, which requires highly efficient likelihood calculations especially for complex pedigrees. Algorithms for likelihood calculations for pedigree data are reviewed in chapter 2. Besides exact likelihood calculation methods and MCMC, a Sequential Importance Sampling (SIS) approach has been proposed to enable calculations for large pedigrees with large numbers of markers. However, when the system gets large, the variance of the importance sampling weights increases while both efficiency and accuracy of the method decrease. We propose an optimization algorithm for calculating the likelihood of general pedigrees in Chapter 3. We incorporate a resampling strategy into SIS to reduce the variance inflation problem. A successful linkage analysis may identify a linkage region of interest containing hundreds of genes at a magnitude of perhaps ten to thirty centiMorgans. A follow-up association (or so-called linkage disequilibrium) analysis can

provide much finer gene-mapping but is subject to greater multiple testing problems. In Chapter 4, we present a method for determining whether an association result is responsible for a non-parametric linkage result for binary traits in general pedigrees. The correlation between family frequency of a variant of interest and family LOD score is used as a measure of whether the association between a given variant at a marker and the disease status can help to explain a significant linkage result seen in the collection of families in the region around the marker.

Disease Gene Mapping in General Pedigrees

by

Li Li

A dissertation submitted to the Graduate Faculty of

North Carolina State University

in partial fulfillment of the

requirements for the Degree of

Doctor of Philosophy

BIOINFORMATICS

Raleigh

2004

APPROVED BY:

SHARON R. BROWNING	MARGARET G. EHM
ZHAO-BANG ZENG	MICHAEL D. PURUGGANAN

BRUCE S. WEIR
CHAIR OF ADVISORY COMMITTEE

To my husband, my parents and my parents-in-law

Biography

Li Li was born in 1976, in Jingzhou, Hubei Province, the People's Republic of China. In 1994, she entered the School of Life Sciences at Wuhan University in China, where she received her B.S. in Biology in June 1998. Li Li was enrolled in Bioinformatics Program at North Carolina State University in 2000. She worked with Dr. Sharon R. Browning on optimizing algorithms for likelihood calculations on large pedigrees. In May 2003, she received her M.S in Bioinformatics. Since then, Li Li has been studying for her Ph.D. in the same program at NCSU, and working as a graduate industrial trainee in Genetic Data Sciences group at GlaxoSmithKline. One of her main research focuses is methodology development for disease gene mapping using general pedigrees. Li Li is also interested in other population or nuclear family based disease gene mapping methods.

Acknowledgements

I would like to express my deepest gratitude to Dr. Bruce S. Weir and Dr. Sharon. R. Browning for being my advisors at school and at work. I greatly appreciate them, especially Dr. Browning, for their guidance and encouragement throughout my graduate studies.

I would like to thank the members of my advisory committee for their helpful advice and support. They have been a wonderful committee that I've benefited a lot from them.

I would like to thank all the members in Genetic Data Sciences group at GSK in US for their help and friendship. I would like to express my deepest appreciation to Dr. Meg Ehm for giving me the opportunity to work in the group and guiding me throughout work. Special thanks for Dr. Dmitri Zaykin for his helpful discussions.

I would also like to thank Dr. Barbara Sherry, Juliebeth Briseno and Debra Hibbard for their support and paperwork.

Finally, I thank my husband, Kejun Liu, for his love, wonderful support and helpful discussions.

Table of contents

List of tables

List of figures

Chapter 1

Introduction to disease gene mapping

Basic human genetics

A *chromosome* is a double-strand DNA in helical structure. A normal human cell consists of 46 chromosomes that are grouped in pairs: twenty two pairs of *autosomes* and one pair of sex chromosomes. For a pair of the autosomes, both chromosomes are homologous with one inherited from the father and the other from the mother of the individual. The sex chromosomes are denoted as X and Y chromosomes. A female has a pair of X chromosomes derived from both parents, while a male contains one X from the mother and one Y from the father. Sex-linked chromosomes are ignored in the following chapters. It is not difficult to extend methods to be applicable to sex chromosomes with modifications.

The total length of the DNA in a normal human cell is about 3 billion base pairs. A *gene* is a stretch of DNA sequence that codes for some functional proteins. A *locus* has a broader definition in that it can be any position on a chromosome, such as a gene, a segment of DNA or even a single base. Loci showing variation are useful for human genetics analysis. Such loci are called genetic *markers*, which are normally easy to assay. The DNA sequence at a marker may take different forms. We call each variation as an *allele*. If a marker shows two or more alleles in a population, it is said to be *polymorphic*. There are markers – some more polymorphic than others. RFLPs or restriction fragment length polymorphisms (Botstein et al. 1980) were the first genetic markers used. They often have two alleles. Several important disease genes, including those for Huntington's disease and cystic fibrosis, were mapped by RFLPs (Gusella, 1986). Then simple sequence repeats (SSR) or microsatellite markers became available. They are more

polymorphic making them useful for disease gene mapping. Technology advances have made SNPs or single –nucleotide polymorphisms more cost effective to assay. Most of them are diallellic and less informative. However, they are abundant and densely located over chromosomes with a density of 1 SNP per 500~1000 bp. According to the latest release of NCBI dbSNP build 123, ~5M validated SNPs are public available in the database. SNPs are rapidly becoming the marker of choice.

Humans are diploid in that they have two alleles at a single marker. The allele pair forms the *genotype* at the locus. If both alleles of a genotype are of the same type, the individual is *homozygous* for the marker. Otherwise, the individual is *heterozygous*. Normally, the observed genotype data is an *unordered* allele pair or *phase unknown* genotype, which means the parental origin of each allele is unknown. Otherwise the genotype is called phase known. Considering multiple loci across a chromosome, the set of alleles on a chromosome originated from the same parent is called a *haplotype*. An individual's multilocus genotype consists of two haplotypes.

Given a marker with k alleles, it can form $k(k+1)/2$ possible types of phase unknown genotypes. The frequency of an allele or genotype is the proportion of the allele or genotype in a population. If the genotype frequencies in a population at a given autosomal locus can be determined by independent allele frequencies, the population is said to be in HWE or Hardy-Weinberg equilibrium at the locus (Hartl, 1988). In this case, the frequency of a homozygous genotype equals to the square of the frequency of the homozygous allele and the frequency for a heterozygous genotype is twice the product of

frequencies of the two consisting alleles. HWE can be reached after one generation of random mating. Non-random mating can cause deviation from HWE.

Different genotypes at a functional site may result in different characteristics or *phenotypes*, like height or hair color. Phenotypes can be discrete or quantitative. For a disease trait, an individual may be identified as affected or unaffected, or his phenotype can be recorded as some continuous value such as body mass index (BMI) and blood pressure. (Sometimes it might be confusing that the observed genotypes over markers are called phenotypes too in the literature to differentiate them from unobserved phase known data. Others use the term "marker phenotypes".) In this thesis, only binary traits will be considered. The conditional probability of a phenotype given a genotype at the trait locus is the *penetrance* of the genotype. Suppose a functional locus has 2 alleles **A** and **a** with the penetrance of **AA** greater than that of aa. If the penetrance of genotype **AA** is the same as that of **Aa**, we say allele **A** is *dominant* to **a** or the trait is of a dominant genetic model. If only genotype **AA** has relative larger penetrance while Aa and aa are on the same scale, allele **A** is said to be *recessive* or the genetic model of the trait is called recessive. If the penetrance of the heterozygous genotype **Aa** is the average of the other 2 homozygous genotypes, we say the effect of alleles are *additive* or the trait model is additive.

Mendel's first law states an individual has one maternal and one paternal copy of a gene, and a copy of a randomly selected one of the two is passed to an offspring. This random process of passing a copy of a gene to a child is *Mendelian segregation*. During

segregation, gametes are formed by a process called *meiosis*. *Crossover* or *recombination* events may happen on the two homologous chromosomes before the formation of gametes. This results in a phenomenon that a chromosome in a gamete consists of a mixture of maternal and paternal originated DNA. If the gamete mates with another egg or sperm, this chromosome will become one of the two haplotypes of the child.

The specific order and spacing of the markers produces a physical map of them. Besides the physical map of markers on a chromosome, the genetic map is often used in human genetics. These two maps do not have an explicit relationship to each other. Haldane (1919) defined the genetic distance (in Morgans) between any two loci as the expected number of recombination events between them on a gamete. One Morgan indicates the expected number of recombination events between two loci on a gamete is one. The genetic map distance and recombination rates are interchangeable by applying some map functions. Haldane's genetic map function assumes no genetic inference between disjoint chromosome intervals. Kosambi (1944) derived his map function by making assumptions on marginal interference. Both map functions are widely used.

Two loci are said to be *unlinked* if the *recombination rate* between them is 0.5, or *linked* otherwise. Mendel's second law on independent segregations between genes only applies to unlinked loci that are far away from each other on a same chromosome or on different chromosomes. For linked loci, segregations are correlated, which is the basis for disease gene mapping. If two loci are close to each other, the alleles from the same parent tend to

pass to offspring together as recombination events are very unlikely to happen. Hence, *linkage* can be detected by analysis of the extent to which the same parental alleles at two loci cosegregate in pedigrees. Different families may cosegregate different allele combinations at the two loci.

Association refers to dependence of specific alleles at different loci. Assume two diallelic markers with allele A and a, B and b respectively. Let p and q be the frequency of allele A and B respectively, and assume that P_{11} is the frequency of the haplotype consisted of allele A and B at these two loci, the linkage disequilibrium (LD) coefficient $D = P_{11} - pq$ is a measure of the population association between alleles at the two loci. Recombination diminishes the association between loci over generations. The decay speed for LD is determined by a factor of one minus the recombination rate per generation. If a recombination rate between two loci is small, as for tightly linked loci, the LD may be maintained for a long time over generations. The LD decays very fast for unlinked or less correlated loci. Considering a disease variant arises at a point on a specific haplotype, this disease variant is in complete LD with all alleles on the haplotype for all loci. After many generations, the association between the disease variant and the markers in greatest distance decreases most dramatically as a consequence of more frequent recombination events. Only those markers in close proximity maintain a high LD to the disease variant. Such markers are associated with the disease phenotype due to linkage and LD to the disease variant, which provide evidence of linkage/association to the disease locus and may be used to locate the disease locus.

IBD and kinship coefficient between relatives

Individuals are *relatives* if they share some common ancestors over a limited time span. Relatives are genetically similar to each other than to unrelated individuals, because they share alleles from common ancestors. The same copy of an allele inherited from a common ancestor is *Identical by Descent* (IBD). Identical allele type is a necessary but not sufficient condition for IBD. Alleles with the same type but originating from different ancestors are said to be *Identical by State* (IBS). A pair of non-inbred individuals may carry 0, 1, 2 alleles IBD with specific probabilities determined by their relationship. Let (k_0, k_1, k_2) be the set of such probabilities with the subscript indicating the number of alleles shared IBD. These probabilities are called prior IBD probabilities if they are determined solely by relationship without genotype data. Assuming no inbreeding (no marriage between relatives), the prior probabilities are (1, 0, 0) for an unrelated pair, (0, 1, 0) for a parent-offspring pair, (0, 0, 1) for monozygous twins and (.25, .5, .25) for a sib pair. We can also compute the posterior IBD probabilities by taking data into account as well. Methods have been proposed and improved to calculate posterior IBD probabilities (Haseman and Elston, 1972; Amos et al., 1990; Kruglyak et al., 1995; Kruglyak and Lander, 1995).

The *kinship coefficient* is a simple probabilistic measure of IBD between 2 individuals. It will be used in Chapter 4. It is defined as the chance of a randomly chosen allele from first individual being IBD to a randomly chosen allele from the second individual. The relationship between the kinship coefficient and IBD probabilities for a pair of non-inbreeding individuals is $kinship = k_2/2 + k_1/4$. The corresponding prior kinship

coefficients for unrelated, parent-offspring, monozygous twins and sib pairs are 0, 0.25, 0.5 and 0.25. As we can see, different relationships may have the same kinship coefficients but their IBD probabilities are specific to each specific relationship assuming no inbreeding.

Model-based linkage methods

Linkage analysis can be used to map the position of a disease gene or estimate genetic maps. Model-based or parametric linkage methods rely on likelihood approaches. Here, "model", which needs to be specified in advance for the analysis, refers to the mode of inheritance of a trait that is under study. The deviation of recombination fraction between the trait locus and a marker from 0.5 is an indication for the existence of a linkage for a two-point analysis. The LOD score method was first introduced by Morton (1955). The likelihood ratio for linkage at a given recombination fraction to that for no linkage is calculated. The logarithm of the likelihood ratio to base 10 is the LOD score. The maximized value of the LOD score serves as a statistic for the linkage test. Morton (1955) suggested that a LOD score of 3 corresponds to a type I error of 0.001. Many people use it as a threshold for determining significant linkage result. Multipoint linkage analysis is an extension of two-point linkage analysis but with greater power. When a marker map is known, the location score is computed for each hypothetical location of the trait locus relative to the null hypothesis (no linkage). Use of multiple marker loci combines information from markers that are informative in different segregations of the pedigree. But correct map position is crucial for the analysis.

Nonparametric linkage methods

Model-free or nonparametric linkage methods have been used for the study of complex disease as the trait parameters are usually unknown. These methods are based on the correlation between concordant relatives with respect to a trait and the similarity at markers. If the marker is linked to the trait locus, the similarity at the marker of the (phenotypic) concordant relatives should be high as they tend to inherit similar alleles at both marker and trait loci.

Affected sib-pair methods

Affected sib-pair methods detect linkage by testing for excess IBD sharing in affected sib-pairs. As we discussed in the section for IBD, the expected IBD probabilities for 0, 1, 2 IBD sharing in a sib-pair are .25, .5 and 0.25. Counting the number of sib-pairs sharing different IBDs and testing the deviation from expected values by a χ^2 test is an intuitive and simple approach (Cudworth and Woodrow, 1975). A mean test compares the average number of alleles shared by affected sib-pairs with the expected value. A proportion test is based on the proportion of affected sib-pairs sharing 2 alleles IBD. Although the mean test was thought to be generally more powerful than the counting and proportion tests (Blackwelder and Elston, 1985; Knapp, 1994), the true underlying genetic model will affect its performance and may result in the reverse conclusion. Approaches that combine the mean and proportion test have been proposed (Schaid and Nick, 1990; Feingold and Siegmund, 1997). Uncertainty of IBD sharing can be taken into account by considering all possible configurations and weighting them accordingly. Alternatively, tests based on

IBS, which is observable, were proposed (Lange, 1986a and 1986b). Liu and Weir (2004) extend Lange's IBS tests to deal with inbreeding and relatedness in the population. However, IBS based methods are less powerful than IBD based methods, especially when parental data are available (Davis and Weeks, 1997).

Risch (1989, 1990) proposed a likelihood ratio test based on IBD sharing. The likelihood of the observed data is maximized with respect to IBD probabilities. A likelihood ratio test is performed by taking the ratio of the maximized likelihood over the likelihood at the expected IBD probabilities. Improvements in the performance of the test can be achieved by putting constraints to the estimates of IBD probabilities (Holmans, 1993).

In the context of quantitative traits, Haseman and Elston (1972) regresses the squared sib-pair trait difference on the estimated proportion of alleles that are IBD for the sibs at a marker. Whenever a linked trait locus exists, the slope is expected to be negative. Although the method was developed for continuous trait, it is valid for binary traits too.

IBD sharing methods in extended pedigrees

In the case that there are more than a pair of affected sibs in each family, methods are proposed to consider all possible affected pairs in each family or to consider IBD sharing in the whole set of affected individuals per family. For the first type of approach, weights are assigned to each family to account for dependency between pairs from the same family and different family sizes (Sham et al., 1997).

For extended pedigrees, Whittemore and Halpern (1994) introduced two score functions, S_{pairs} and S_{all}. The former is defined to be the number of pairs of alleles from different affected pedigree members that are IBD at the locus. The latter considers IBD sharing in larger sets of affected relatives by putting extra weight on three or more affecteds sharing the same allele IBD. Kruglyak et al. (1996) normalizes the score functions to a standard z-score. Both score functions have been implemented in the GENEHUNTER package. In order to combine scores from different pedigree, weighting factors are used. The optimal choice of scoring function and weighting factors will depend on the underlying disease model. Generally, a z-score test using S_{pairs} performs better in a variety of disease models (McPeek, 1999; Davis and Weeks, 1997). The z-score test was found to be conservative, especially when IBD information is incomplete. Davis et al. (1996) proposed another approach for extended pedigrees. He assigned a score to each affected pair based on the probability that the two share a specific allele IBD in extended pedigrees. The family score is a summation of scores over all possible affected pairs for each pedigree. An empirical *p*-value for the sum of the scores over all pedigrees can be estimated using conditional simulation. A problem with the method lies in its poor performance in dealing with sibships without genotyped parents (Davis and Weeks, 1997).

Model-based vs. Model-free linkage methods

One main drawback of the classical model-based linkage method is that the genetic model for the trait needs to be specified. When the mode of inheritance of a disease is well established, the parametric lod score linkage analysis is undoubtedly a good choice for detecting disease susceptibility genes, because likelihood ratio tests will usually yield

more power compared to model-free tests. However, if the model is incorrectly specified, a conservative lod scores may be produced. Even in the case of Mendelian disease, assuming a dominant model for a true recessive model reduces power dramatically (Risch and Giuffra, 1992; Clerget-Darpoux et al., 1986). It is recommended that model-based linkage analysis should be carried out under multiple disease models. For a complex trait, where multiple genes of small effects may be involved, a disease model with a single major locus is not appropriate in most cases.

It has been shown that a lod score analysis for an assumed recessive mode of inheritance, irrespective of the true mode of the disease, is equivalent to the mean test (Knapp et al., 1994b). But it is not the case for other modes of inheritance. Goldin and Weeks (1993) compared the model-based method to several nonparametric methods and showed the nonparametric linkage methods had lower power than the lod score method under certain conditions. Elston (1998) suggested model-free methods should be used initially to explore underlying mode of inheritance. More powerful model-based methods should be used in the second step when sufficient information has been gained.

The model-based linkage methods based on likelihood have the strength that unknown parameters, such as allele frequencies and recombination fractions, can be estimated. However, it might be a disadvantage for a complex trait as the recombination fraction between the disease locus and a nearby marker tends to be overestimated (Schork et al., 1993; Risch and Giuffra, 1992). IBD sharing probabilities are functions of parameters of the genetic model. It is possible to use estimated IBD probabilities from a non-parametric

method to estimate the parameters as well. However, due to the low degrees of freedom, at most two of parameters can be estimated uniquely.

Association analysis

Linkage studies have a low resolution for disease gene mapping. This is due to the fact that insufficient recombination events happen in each pedigree. A successful linkage analysis may identify a linkage region of interest of perhaps ten to thirty centiMorgans. Association studies are considered an important complement to linkage analysis. An association (or so-called linkage disequilibrium) analysis can identify markers within 1MB of the disease gene. Genotyping technology advances are making genome wide association studies more feasible. Analysis interpretation of theses results are more difficult. With the increasing amount of high density SNPs being available, genome wide scans instead of candidate gene studies are widely used in recent years. Association studies, however, is unfortunately subject to greater multiple testing problems.

Population-based association methods

The classical case control study compares the difference in allele or genotype frequencies in a sample of cases and a sample of controls. These methods test the hypothesis that an allele or genotype is correlated with affection status. In the simplest case of a diallelic locus, a simple χ^2 test of a 2×2 contingency table can be performed. Terwilliger (1995) constructed a likelihood-ratio test with one degree of freedom to test for linkage disequilibrium, which can be applied to multiallelic marker systems and extended to multiple marker loci simultaneously. It maintains higher power than the conventional

case control test. If multiple affected individuals are available from a family but with population controls, population based methods have been extended to make use of all data by taking correlations of related cases into account. They have been proved to be more powerful than methods using only one case per family (Browning et al., 2004; Risch and Teng, 1998).

Association detected by such population-based designs can be a result of the existence of LD between the tested allele and the disease-causing variant, which is what we are interested in. However, these association methods have some drawbacks that it can also detect spurious association due to population structure. Population structure means the population under study comes from genetically different groups, for example, unmatched cases and controls or non-randomly mating in a population over several generations (Pericak-Vance, 1998). One way to solve this problem is to use different study designs, such as using family based methods instead. Alternatively, we can assume the stratification effect is the same for the whole genome and estimate the effects from background markers. The genomic control method includes a set of markers independent from the trait of interest to determine the degree to which case control association findings are inflated due to population structure and corrects the bias in the association analysis (Devlin and Roeder, 1999; Bacanu et al., 2000). Although the Genomic Control method often performs well, it may be anticonservative if too few loci are used or overcorrect when too many loci are used which results in a loss of power (Marchini et al., 2004). Pritchard et al. (1999 and 2000) uses unlinked markers to identify the presence of population stratification and assign samples to subpopulations to do the association test.

Family-based association methods

Family-based association tests, which use parental alleles not transmitted to the affected child as controls, such as the HRR test (Falk and Rubinsterin, 1987) and TDT test, are not affected by population stratification. Under the null hypothesis of no linkage and association, the transmission of an allele from a parent to an offspring is a random process. Although the original TDT test was proposed as a test for linkage (Spielman et al., 1993), it is more widely used as a test for association for binary markers. As a test of linkage, TDT can accommodate any pedigree structures as the transmissions at the marker locus are independent from each other. In case of testing for association in the presence of linkage, this independence of transmissions from parents to affected relatives does not hold (Martin et al., 1997) and only a single affected sib can be used. The original TDT is designed for diallellic markers. Many modifications have been made to handle multiallelic marker. One easy extension was proposed by Spielman and Ewens (1996). They compare the number of times each marker allele is transmitted to affected offspring to the number of times it is not and summed over all alleles at the marker loci.

The TDT test was initially constructed to test for linkage or association in families with known parental genotypes. It is very difficult or impossible to obtain parental genotype information when the disease has a late age of onset. One way to handle missing parental data is to make use of other unaffected relatives in the family. Curtis (1997) used unaffected siblings as controls to maintain the robustness of the test against bias due to population stratification. A similar method, the S-TDT test, was proposed by Spielman

and Ewens (1998) to deal with missing parental data. It compares the observed number of an allele in affected children with the expected number under null hypothesis, conditioned on the observed distribution of marker genotypes in the whole sibship. A single affected and a single unaffected sibling with different genotypes per family is a requirement for this test to be a valid test for association. The overall test by combining TDT and S-TDT is very useful for data consisting of a mixture of families with known parental genotypes and families with missing parental data. TDT is preferred if the data can be analyzed by either method as the TDT test has more power than S-TDT. The sibship disequilibrium test (SDT) was presented by Horvath and Laird (1998) to handle larger sibships. It requires at least one discordant pair per family. Another way to handle missing parental information is to reconstruct parental genotypes. Bias on inferred parental genotypes arises in the presence of linkage (Curtis, 1997), but it can be corrected by a reconstruction approach developed by Knapp (1999). In contrast to TDT and its various extensions, a likelihood-based test for trios was proposed by Weinberg et al. (1998). This test is based on the joint transmission from pairs of parents instead of one parent at a time. An extended version that uses an Expectation-Maximization procedure allows for missing parental data (Weinberg, 1999).

The TDT test can also be generalized to extended pedigrees. Martin et al. (1997) proposed two test statistics using data from all the affected children. They are more powerful than the normal TDT using a single affected child. A pedigree disequilibrium test (PDT) was developed by Martin et al. (2000). It can analyze linkage disequilibrium in pedigrees of any size. It defines a summary statistic for a pedigree by combining

scores over all trios and discordant siblings in the pedigree. Weights accounting for pedigree sizes were incorporated later (Martin et al., 2001).

Association tests for multiple markers

A MANOVA approach was presented by Xiong et al. (2002) to test multiple markers simultaneously in a case-control design. The dependent structure between markers is incorporated into the test. Their analytical and simulation studies showed the test has a higher power than the classic chi-square test.

The TDT test has also been extended to handle more than one marker locus. Based on the assumption that the disease locus is located between two marker loci and the parental haplotypes are known, Wilson (1997) studied the marker transmission of two multiallelic marker loci from parents to affected offspring. But it applies only to phase-known data. Another method analyzing multiple tightly linked markers simultaneously by Zhao et al. (2000) has been shown to be more powerful than other existing methods.

Several recent association approaches are based on haplotype sharing rather than single marker sharing. The underlying principle is that cases are expected to share a haplotype sequence surrounding the disease-causing variant. In unrelated individuals, conserved ancestral chromosome segments usually extend 1-2cM (March, 1999). Clayton and Jones (1999) generalized the TDT test to detect association of haplotypes by comparing the transmitted haplotypes with untransmitted haplotypes and aimed to detect regions of linkage disequilibrium where the susceptibility gene is located. Clayton (1999) proposed a likelihood-based test to handle phase-unknown data and implemented it in TRANSMIT.

The mosaic structure of LD along chromosomes has been utilized recently. Sites of closely located SNPs which are inherited in blocks are called as haplotype blocks with only a few common haplotypes which account for a large proportion of chromosomes. Daly et al. (2001) developed a haplotype block approach, in which they performed LD mapping based on haplotype blocks.

Conclusions

Disease gene mapping is one of the main focuses in human genetics. The existence of linkage disequilibrium between a disease variant and a marker allele or haplotype enables us to locate the disease gene through the analysis of genetic markers. Linkage methods analyze the cosegregation pattern of alleles at a marker and the disease locus. The model-based linkage methods are most powerful for Mendelian diseases with known disease model. The IBD sharing based linkage methods have the advantage of no assumptions on disease models, which fit the situation of complex diseases well. However, the resolution for disease gene mapping of both types of linkage methods is low. Conversely, association methods may locate a gene in a much finer map. The allele-specific association at a marker and the disease locus can be detected by comparing the difference of marker allele frequencies in cases and controls. Population-based methods use independent individuals as cases and controls, while TDT-based methods use transmitted and untransmitted alleles from a parent to offspring as cases and controls. Various methods have been developed for single-point, multi-point or haplotype based tests.

References

Amos, C., Dawson, D. V. and Elston, R. C. (1990) The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. *Am. J. Hum. Genet.* 47, 842-853

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

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

Browning, S.R. (2004) Case-control single-marker and haplotypic association analysis of pedigree data. *Genet Epidemiol. (in press)*

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

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

Clayton, D. (1999) A generalization of the transmission/disequilibrium tests for uncertain- haplotypes transmission. *Am. J. Hum. Genet.* 65:1170-1177.

Cudworth, A. G. and Woodrew, J. C. (1975) Evidence for HLA-linked genes in 'juvenile' diabetes mellitus. *Brit. Med. J.* 3, 133-135

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

Daly, M. J., Rioux, J. D. et al. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* 29:229-232

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

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

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

Elston, R. C. (1998) Linkage and association. *Genet. Epidemiol.* 15, 565-576

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

Feingold, E. and Siegmund, D. O. (1997) Strategies for mapping heterozygous recessive traits by allele-sharing methods. *Am. J. Hum. Genet*. 60, 965-978

Goldin, L. R. and Weeks, D. E. Two locus models of disease: comparison of likelihood and nonparametric linkage methods. *Am J Hum Genet* 53:908-915, 1993.

Gusella, J. F. (1986), DNA polymorphism and human disease. *Ann. Rev. Biochem.* 55, 831-854

Haldane, J. B. S. (1919) The Combination of linkage values and the calculations of distances between the loci of linked factors. *J. Genet.* 8: 299 – 309

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

Hartl, D. L. (1988), *A Primer of Human Genetics*. Sunderland, Mass.: Sinauer Associates.

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

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

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

Knapp, M., Seuchter, S. and Baur, M. (1994) Linkage analysis in nuclear families. 1. Optimality criteria for affected sib-pair tests. *Hum. Hered.* 44, 37-43

Knapp, M., Seuchter, S. and Baur, M. (1994b) Linkage analysis in nuclear families. II. Relationship between affected sib-pair tests and lod score analysis. *Hum. Hered.* 44, 44-51

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

Kosambi, D. D. (1944) The estimation of map distances from recombination values. *Ann. Eugen.* 12: 172 – 75

Kruglyak, L., Daly, M. J. and Lander, E. S. (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping, *Am. J. Hum. Genet.* 56, 519-527

Kruglyak, L. and Lander, E. S. (1995) Complete multipoint sib pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* 57, 439-454

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

Lange, K. (1986a) A test statistic for the affected-sib-set method. *Ann. Hum. Genet.* 50, 283-290

Lange, K. (1986b) The affected sib-pair method using identity by descent relations. *Am. J. Hum. Genet.* 39, 148-150

Liu, W. and Weir, B. S. (2004) Affected sib pair tests in inbred populations. *Ann. Hum. Genet.* Published online in Sep 2003.

March, R. E. (1999). Gene mapping by linkage and association analysis. *Mol. Biotechnol.* 13, 113-122.

Marchini, J., Cardon, L.R., Philips, M. S. and Donnelly, P (2004) The effect of human population structure on large genetic association studies. *Nat. Genet.* 36(5):512-517

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

Martin, E. R., Monks, S. A., Warren, L. L. and Norman, N. L (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am. J. Hum. Genet*. 67, 146-157

Martin, E. R., Bass, M. P. and Norman, N. L (2000) correcting for a potential bias in the pedigree disequilibrium test. *Am. J. Hum. Genet*. 68, 1065-1067

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

Morton, N. E. and MacLean, C. J. (1955) Analysis of family resemblance. III. Complex segregation of quantitative traits. *Am. J. Hum. Genet.* 26, 489-503

Ott, J. (1992), *Analysis of Human Genetics Linkage,* 2nd. Ed., Johns Hopkins University Press, Baltimore

Pericak-Vance, M. A. (1998) Linkage disequilibrium and allelic association. In: Haines, J. L., Pericak-Vance, M. A, editors. Approaches to gene mapping in complex human disease. New York, Willey-Liss. pp 323-333

Pritchard, J. K. and Rosenberg, N. A. (1999) Use of unlined genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* 65, 220-228

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67,170-181

Rao, D. C. and Province, M. A. (2001), Advances in Genetics (42): Genetic Dissection of Complex Traits, Academic Press

Risch, N. (1989) Genetics of IDDM: Evidence for complex inheritance with HLA. *Genet. Epidemiol.* 6, 143-148

Risch, N. (1990) 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. and Giuffra, L. (1992) Model misspecification and multipoint linkage analysis. Human Heredity 42, 77-92

Risch, N. and Teng, J. (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex disease. I. DAN pooling. Genome Research, 8, 1273-1288

Schaid, D. J. and Nick, T. G. (1990) Sib-pair linkage tests for disease susceptibility loci: common tests vs. the asymptotically most powerful test. *Genet. Epidemiol.* 7, 359-370

Sham, P. C., Zhao, J. H. and Curtis, D. (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann. Hum. Genet.* 61, 61-69

Schork, N. J., Boehnke, M., Terwilliger, J.D. and Ott, J. (1993) Two-trait-locus linkage analysis: A powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet*. 53, 1127-1136

Spielman, R. S. and Ewens, W. J. (1996) The TDT and other family-based tests for linkage disequilibrium and association . *Am. J. Hum. Genet*. 59, 983-989

Spielman, R. S. and Ewens, W. J. (1998) A sibship test in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet*. 62, 450-458

Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet*. 52, 506-116

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

Weinberg, C. R., Wilcox, A. J. and Lie, R. T. (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 may be subject to parental imprinting. *Am. J. Hum. Genet.* 62, 969-978

Weinberg, C. R. (1999) Allowing for missing parents in genetic studies of case-parent-triads. *Am. J. Hum. Genet.* 64, 1186-1193

Xiong, M., Zhao, J. and Boerwinkle, E. (2002) Generalized T^2 test for Genome Association Studies. *Am. J. Hum. Genet.* 70, 1258-1268

Zhao, H., Zhang, S et al. (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am. J. Hum. Genet.* 67, 936-946

Chapter 2

Likelihood on general pedigrees

Abstract

Probabilities of genotypic data from multiple linked marker loci for related individuals are used as likelihoods of gene locations for gene-mapping, or as likelihood of other parameters of interest in human genetics. These probabilities may be calculated exactly with the Elston-Stewart algorithm (Elston and Stewart, 1971) or Lander-Green algorithm (Lander and Green, 1987). Computational time for the Elston-Stewart algorithm scales linearly with the number of meioses in a pedigree and exponentially with the number of marker loci on a chromosome, while the Lander-Green algorithm scales linearly with the number of markers but exponentially with the number of loci. With the recent development in genetics and molecular biology techniques, large-scale marker data has become available, which requires highly efficient algorithms especially for complex pedigrees. Monte Carlo approximation approaches have been developed as a good alternative, which scale linearly with both marker number and pedigree size.

Definitions and notations

A pedigree is a specification of genealogical structure among a set of individuals (Thompson, 2000). It consists of units of nuclear families (parents and offspring). *Founders* are individuals with no parents in a pedigree. All other individuals are *non-founders*. Normally founders are assumed to be independent for human pedigrees. Figure 2.1 lists some graphic representations of a simple three-generation pedigree. Squares refer to males and circles to females. In figure 2.1a and figures in chapter 3, no lines connect a married couple. Each meiosis event is represented by a line from a parent to an offspring. We index meioses by numbers as shown in figure 2.1a. In figure 2.1b and figures in Chapter 4 (figure 4.1), pedigrees are drawn differently in that two individuals are connected by a horizontal line below them if they are married and sibs share a single or both parents are connected by a horizontal line above them. A single vertical line is used to connect the parents and the set of sibs. Nuclear family information is very clear by this type of presentation. These two types of graphic presentations of pedigree structure will be used for different types of likelihood calculations later. Shadings for squares and circles may be used to represent information in different contexts. For example, in chapter 3, shaded shapes indicates individuals of interest to infer relationships, while in chapter 4, they mean affected individuals. A *path* along a pedigree is defined as a specific sequence starting from an individual connected by direct relatives (parents, spouse or offspring) to another individual. For the example in figure 2.1, one possible path from A to B is AC (E or F) DB.

We define an inheritance indicator as 0 if the offspring obtains a copy of the parent's maternal allele, or 1 for copy of the parent's paternal allele in meiosis. Each value has an equal probability of 0.5 assuming a random segregation. For figure 2.1a, the six meiosis can be represented by a vector of inheritance indicators, v, of length 6, such as (0, 0, 1, 0, 0, 0). As each meiosis event is independent from each other, the total number of possible values of an inheritance vector of length 6 is 2^6.

The observed data, denoted as X, for a pedigree may be phenotypes of a specific trait or genotypes over markers (normally phase unknown) for some family members. For an observed data set, we are interested in making inference for some parameters of interest. Such parameters of interest are denoted as θ. θ may refer to pairwise relationship as in chapter 3 rather than the location of a disease locus as in a linkage analysis. We denote a latent variable such that it determines the observed data. A latent variable is a type of missing data and is different from parameters in that they are not of primary interest. The latent variable can be the unobserved ordered genotypes over all members in the pedigree. We use notation G in this case. Alternatively it can be the underlying inheritance vectors for the pedigree, denoted as V. We use S to represent the latent variable in general. The likelihood of the parameter of interest for a pedigree is

$$\begin{aligned} L(\theta \mid X) &= P(X) \\ &= \sum\nolimits_S P_\theta(X, S) \\ &= \sum\nolimits_S P_\theta(X \mid S) P_\theta(S) \end{aligned}$$

The probability of an observed phenotype given a specific value of a latent variable is the *emission probability* of the latent variable at the value. If the phenotype is for a trait of interest and the latent variable is a genotype, it is just the penetrance of the genotype as defined in the previous chapter. If the observed phenotype is the genotype data itself, the emission probability takes value 1 if it is compatible with the given latent variable or takes value 0 otherwise. The probability of a founder's genotype is the population frequency of the genotype. The probability of a non-founder's genotype is the segregation probability of the genotype conditional on his/her parents' genotypes. If an individual whose parents are Aa and Aa, has genotype AA, the segregation probability of the individual given the parents' genotypes is .25. The segregation probabilities for Aa and aa offspring are 0.5 and 0.25.

Pedigrees are assumed to be independent in each sample. The likelihood for multiple pedigrees is the product of likelihoods for individual pedigrees. In the following discussion, we will focus on the likelihood for a single pedigree.

Single point likelihood on pedigrees

For simplicity, we first consider a single locus. There are several methods for the calculation of the exact likelihood. Different methods are suitable for pedigree structure of different complexity.

Exact likelihood calculation for small pedigrees

One possible method for single point likelihood calculation for a pedigree is to write down all possible genotype combinations of all members and sum over them (Thompson, 1986). The probability of data given a specific genotype combination is the product of the probabilities of founder genotypes, the probabilities of offspring genotypes given parental genotypes, and the probability of observed phenotypes given genotypes over individuals. This enumeration method is applicable only for very simple pedigree structures as the number of possible configurations increases exponentially with the family size. The base for the exponential function is the number of possible genotypes at the locus. For the pedigree in figure 2.1, the total number of possible genotype combinations is $3^6 = 729$ for a diallelic locus.

Another possible method is to use a recursive algorithm to obtain the IBD probabilities in the observed individuals and compute the likelihood (Thompson, 1986). For example, if two individuals have genotypes AA, Aa at a locus and they share 1 allele IBD, then the probability of the observed genotype is $p^2(1-p)$, where p is the frequency for A at the diallelic locus. Again, it is very hard to extend this method to larger pedigrees. When a large number of individuals are available, either very long computational time (if no prior IBD calculations are stored) or large memory capability (if prior IBD calculations are stored) is required.

"Peeling" algorithm for large pedigrees

In order to make calculations feasible for large pedigrees, several authors laid the foundations of an approach that has been widely used since 1970s (Hilden, 1970; Elston and Stewart, 1971; Heuch and Li, 1972). This so-called "peeling" (Cannings et al, 1978) or Elston-Stewart algorithm is a sequential summation of probabilities along a pedigree.

A first-order Hidden Markov Chain has the property that given a hidden state at the current position, the hidden states before and after the position are independent of each other. "Hidden" means the state at a position is unknown. A pedigree can form such a structure: given the genotype for an individual, the observed data of the individual, the genotypes of his/her parents and of his/her offspring are independent from each other. The genotypes are hidden states. By using this property, the "peeling" algorithm limits the number of individuals whose genotypes must be considered jointly at each step. An individual is called a *pivot* if he/she connects a processed nuclear family that he/she is in to the rest of the unprocessed pedigree. Starting at the edges of a pedigree, either founders or offspring in the last generation, summing over non-pivot members of the nuclear family results in a function depending only on the pivot's genotype. If a pivot is parental in the nuclear family, the probability of observed data conditional on the pivot's genotype is calculated. In case of the pivot being an offspring in the nuclear family, the joint probability of observed data and the pivot's genotype is obtained. Then we can move to the next nuclear family that is connected to the processed one by the pivot and repeat the process. The movement can be upward (from offspring to parents) or downward (from parents to offspring) in a pedigree. This process continues sequentially

until all nuclear families in a pedigree have been processed. There may be multiple moving sequences, i.e. sequences of "peeling" for a pedigree.

For the pedigree example in figure 2.1b, there are 3 nuclear families connected to each other by pivots C and D. Figure 2.2 showed one possible peeling sequence. We can start from nuclear family consisted of (A, C) and move to the second family (C, E, F, D), and finally arrive at family (D, B). Let X_j and G_j denote the observed data and unknown genotype data for individual j, where $j = A, B, C, D, E, F$. Detailed calculations for the example are as follows:

Step 1. For all possible values of g, calculate $P_\theta(X_A \mid G_C = g)$

Step 2. Calculate $P_\theta(X_A, X_C, X_E, X_F, G_D = g)$ by calculating $P_\theta(X_A, X_C \mid G_E, G_F)$ first.

Step 3. Calculate $P_\theta(X_A, X_C, X_E, X_F, X_D, G_B = g)$ first and get the likelihood by

$$
\begin{aligned}
&L(\theta \mid data) \\
&= P_\theta(X_A, X_C, X_E, X_F, X_D, X_B) \\
&= \sum_g P_\theta(X_A, X_C, X_E, X_F, X_D, G_B = g) P_\theta(X_B \mid G_B = g)
\end{aligned}
$$

As long as a pedigree is simple (no loops), the "peeling" algorithm is applicable for single point likelihood calculation on pedigrees of any size in theory. The existence of loops in a pedigree dramatically increases the complexity of the algorithm. In case of loops, a single pivot for a nuclear family may not be enough to satisfy the independence relationship required by the first-order Markov chain structure. A set of individuals needs to be identified first to make it feasible. Hence, we need to consider the combination of

the set of individuals simultaneously for the sequential probability calculation. As in the enumeration method introduced previously, the number of possible combinations is an exponential function of the number of individuals in the defined set. Complexity of the algorithm increases rapidly over the number of individuals in the defined set. It is crucial to find a smallest set of such dividing individuals.

Likelihood on pedigrees over multiple loci

If we consider l loci together, the capitalized letters X, G, V or S refers to observed data, unobserved genotype, inheritance vector or latent variable for all individuals at all loci. Let i be the index of a given locus, where $i = 1,...,l$. The subscript for X_i, G_i, V_i means joint data or variables from locus 1 up to i. We also use corresponding small letters to indicate individual values at a specific locus. Let n be the total number of individuals and m be the number of meiosis in a pedigree. Let j be the jth individual, $x_{i,j}$ and $g_{i,j}$ denote the observed data and unobserved ordered genotype for individual $j, j = 1,...,n$ at locus i. The notation for $v_{i,j}$ is slightly different from the above, where i is the index of a given locus but $j, j = 1,..,m$ is the index for a meiosis instead of an individual. $x_{i,.}$, $g_{i,.}$ and $v_{i,.}$ are observed data, unobserved genotypes and inheritance vector for locus i over all individuals or meiosis. On the other hand, $x_{.,j}$ and $g_{.,j}$ indicate observed data and unobserved ordered genotypes for individual j over all loci, where $v_{.,j}$ for inheritance vectors for meiosis j over all loci. Define r_i as the recombination fraction between locus i and $i+1$.

When loci are independent, the segregations over them are independent as indicated by Mendel's second law. We can obtain the likelihood for a pedigree over multiple independent loci simply by taking products of single point likelihoods.

With increasing density of genetic markers, loci are likely to be dependent in human disease gene mapping. The segregations over loci in a pedigree are correlated due to linkage between them. It is not appropriate to take the products of individual likelihoods as the overall likelihood over loci. Instead, the correlation between markers needs to be incorporated into the likelihood calculation. Luckily, we can use some function of the recombination rate between two loci to do so.

Multilocus "Peeling" algorithm

The "Peeling' algorithm can be generalized to multiple loci. Instead of considering genotypes at a single locus, we consider the joint multilocus genotypes together. Again, the multilocus genotype probability of founders equals the population frequency of the genotype. If markers are linked but in linkage equilibrium (LE) with each other, the multilocus genotype frequency in a population is the product of frequencies for each locus. Otherwise, if there is linkage disequilibrium (LD), the joint frequency needs to be estimated from the population. The probability of the multilocus genotype of a non-founder conditional on his/her parents is not a simple countable segregation probability unless genotype data are phase known. This conditional probability depends on recombination rates between markers. Assume three markers with two alleles A and a, B and b, C and c at each locus and recombination frequencies r_{AB}, r_{BC} between the adjacent loci. Assume a pair of parents have ordered genotypes AABbCc (or ABC/Abc with "/"

separating two phase-known haplotypes) and AaBbCc (or ABC/abc). If we know a recombination event occurred between the second and third markers in the meiosis from the first parent and no recombination happened for the other, the probability of an offspring, AABbCC (AbC/ABC) conditional on their parent genotypes and observed recombination event is $\frac{1}{4}(1-r_{AB})^2 r_{BC}(1-r_{BC})$. In reality, the phase of genotypes and recombination events are seldom known. Then we need to sum over all possible phase configurations and recombination events to get the segregation probabilities over multiple dependent loci. Although computation time of the "peeling" algorithm scales linearly with the pedigree size, it grows exponentially with the number of markers. This algorithm is useful only for pedigrees with a small number of markers (~ 8). It has been implemented in the software package LINKAGE (Lathrop et al., 1984).

Lander-Green algorithm

The Lander-Green algorithm (Lander and Green, 1987) is similar to the Elston-Stewart ("peeling") method in that it is based on a first-order Hidden Markov chain as well. Indeed, both algorithms are different applications of Baum's algorithm (Baum, 1972). However, the Markov chain is built along a chromosome for Lander-Green algorithm rather than along a pedigree as in the Elston-Stewart algorithm. The hidden state for this algorithm is the inheritance vector for each locus. Assuming HWE, no LD between loci and no genetic interference, the conditional independence structure provides that, conditional on the inheritance vector ($v_{i,\cdot}$) at locus i, the observed data at the locus and the inheritance vectors ($v_{i-1,\cdot}$ and $v_{i+1,\cdot}$) at the adjacent two loci $i-1$ and $i+1$ are independent of each other. Figure 2.3 shows the Hidden Markov Structure graphically.

The transition probability is $P_\theta(v_{i+1,\cdot} \mid v_{i,\cdot}) = r_i^{|v_{i+1,\cdot} - v_{i,\cdot}|}(1-r_i)^{1-|v_{i+1,\cdot} - v_{i,\cdot}|}$, where $|v_{i+1,\cdot} - v_{i,\cdot}|$ is the number of coordinates that differ between $v_{i+1,\cdot}$ and $v_{i,\cdot}$. The overall likelihood of parameter θ for a pedigree can be rewritten as

$$\begin{aligned} L(\theta \mid X) &= \sum_V P_\theta(X \mid V) P_\theta(V) \\ &= \sum_V \left(\prod_i P_\theta(x_{i,\cdot} \mid v_{i,\cdot}) \right) \left(P_\theta(v_{1,\cdot}) \prod_{i=2\ldots L} P_\theta(v_{i,\cdot} \mid v_{i-1,\cdot}) \right) \end{aligned}$$

Since the possible space for an inheritance vector at one locus is 2^m, the number of possible spaces for V over all loci is 2^{ml}. By the original design of the Lander-Green algorithm that considers combinations of inheritance at adjacent two loci one a time, the computation complexity of the algorithm is $4^m(l-1)$, which scales linearly with marker number but exponentially with meiosis number. This limits its use to small pedigrees with a large number of markers.

Many improvements have been made to reduce the complexity of the algorithm so that more pedigree members can be analyzed. This algorithm has been implemented in two popular linkage packages GENEHUNTER (Kruglyak et al., 1996) and MERLIN (Abecassis et al., 2002). Kruglyak et al. (1995) first reduced the computational complexity to the order of $2^m m(l-1)$ by taking advantage of dependencies in the Markov transition. Later, he made use of the symmetric property of founder phases to reduce the amount of calculations (Kruglyak et al., 1996). Kruglyak and Lander (1998) further sped up calculations by using a discrete Fourier transformation representation. Markianos et al. (2001a and b) made efforts to restrict the available inheritance space by the observed data and combining equivalent vectors together. The linkage package,

Merlin (Abecassis et al., 2002), incorporated the "divide and conquer" algorithm first proposed by Idury and Elston (1997). However, the Lander-Green algorithm is still exponentially scaled with the size of a pedigree despite all the improvements that increase its applicability. Normally, a pedigree with < 20 non-founders can be handled in practice.

Monte Carlo estimation on general pedigrees

Exact likelihood calculation on pedigrees has its limitation in either the number of markers for Elston-Stewart algorithm or the pedigree size for Lander-Green algorithm. When large numbers of individuals and markers are available, Monte Carlo estimation can be used to handle such data. Instead of considering the full possible spaces of the latent variable S (either the unobserved genotype or the inheritance vector), a single realization of S is obtained by a sampling procedure from one sampling distribution. A realization of a latent variable is a complete set of values over all loci and individuals or meiosis. After a large number of independent repetitions of the sampling process, either an unbiased likelihood estimator or estimator for relative likelihood ratio can be obtained. In order to get a good estimator with small variance, the choice of the sampling distribution for S, denoted by $P^*(S)$, and the sampling strategy are very important.

Effective sampling methods normally build upon importance sampling, where the realizations from sampling are weighted in such a way that terms making larger contributions to the overall estimation of likelihood or relative ratio are realized with

larger probabilities (Thompson, 2000). The effectiveness of importance sampling depends on the sampling distribution of the latent variable. A similar shape of the sampling distribution $P^*(S)$ to the joint probability $P_\theta(S, X)$ reduces the variance of the Monte Carlo estimator. A distribution which is close to $P_\theta(S \mid X)$ may be a good alternative. For existing methods in the literature, there are different choices of the sampling distributions for different latent variables and sampling methods, as will be discussed below.

Two major Monte Carlo approaches have been used for pedigree inferences. One is sequential importance sampling (Kong et al., 1994; Iwin et al., 1994; Zachary et al., 2003). For each independent realization, the inheritance vector at each locus is sampled sequentially conditional on the observed data and sampled vectors of previous loci. A weight for the realization is calculated sequentially and simultaneously during the sequential sampling process. Then the average weight over N simulations is the unbiased estimator of the likelihood. The detailed algorithm and one extension we proposed will be covered in chapter 3.

The other major category of existing methods is Markov Chain Monte Carlo (MCMC). A Markov Chain structure is formed from one realization of the latent variable to another. An initial realization of the latent variable is generated from some prior distribution. Then a new realization is proposed from a sampling distribution conditional on the current realization and data. The proposed realization is either accepted and replaces the existing one as the current state, or discarded. Updating or rejection of new proposed realizations

is repeated for a large number of times until the equilibrium distribution of the Markov Chain is reached. Then inference of likelihood can be made. Estimation of the likelihood by MCMC requires knowledge of the distribution of latent variables, which is hard to accomplish. However, a relative likelihood as opposed to the likelihood of a specific parameter value can be estimated instead. The distribution terms cancel out when taking the ratio of the two likelihoods. MCMC based methods scale linearly with both pedigree size and marker number so they can handle hundreds of people with tens of markers

The movement (proposal + acceptance) from states to states is essential for the convergence of the MCMC. Improvements on methods for proposing new states can be achieved by taking some form of importance sampling as discussed before. The probability of accepting or rejecting a proposed new state depends on the different MCMC algorithm used. For Metropolis-Hasting algorithms (Hastings, 1970), such probability is called Hasting ratio. It is the ratio of transition probabilities from the new to the current state and the reverse times the odds ratio of the new state versus the current one. When the transition probabilities from one state to another are reversible, the Hasting ratio reduces to the odds ratio of the two states, and the algorithm becomes the so-called Metropolis algorithm (Metropolis et al., 1953).

Initially, single-site updating methods were proposed such that the latent variable at one unit, i.e. genotype for an individual at a locus or an inheritance indicator at a locus for a meiosis, is updated one at a time. Thompson and Guo (1991) use a Gibbs sampler to update latent genotype variables. The Gibbs sampler samples one variable conditional on

the remaining variables with current values and data. The acceptance probability for a new state is one. All variables are updated sequentially in the same fashion. Alternatively, using inheritance vectors as latent variable and updating one inheritance indicator at a time was presented by Thompson (1994). The Metropolis algorithm is used as the transition probabilities from state "0" of an inheritance indicator to state "1" and the reverse are equal. The first-order Markov Chain structure simplifies the calculation greatly as the proposed new inheritance indicator changes only the recombination events between the locus and the two adjacent loci.

Updating only a single variable in high-dimensional spaces means the proposed changes are small, which is accompanied by higher probabilities for new states to be proposed and accepted as they are more likely to be consistent with the data. On the other hand, small changes result in slow convergence speed. One way to solve the problem is to improve the efficiency of sampler by enhancing the movement around the space. Lin et al. (1993, 1994) use a "temperature" parameter, which is normally used in simulated annealing, in the sampling distribution for new genotypes. Chances are increased for large movements between different parts of the space.

However, when dealing with multiple linked loci, single-site updating methods are still inefficient. Methods updating several variables jointly have been developed to improve the Monte Carlo estimates. Kong (1991) updates all genotypes at a single locus conditional on those at neighboring loci. Jensen et al (1995) updates genotypes of blocks of individuals at several loci together. Many other joint-updating methods update

inheritance vectors instead of genotypes. Figure 2.4 shows the sampling units for four MCMC samplers using inheritance vectors as the latent variable. The last three samplers are joint-updating methods. Based on the idea of Kong (1991), Heath (1997) developed the locus-by-locus sampler (L-sampler). The components of the inheritance vector at a locus $v_{i,\cdot}$ are updated jointly. The sampling distribution of the inheritance vector at a given locus has a very simple form in that it depends only on the observed data at the current locus, and inheritance vectors at two adjacent loci. The L-sampler works well where there are extended ancestral paths of descent in a pedigree, but suffers from poor mixing due to tightly linked loci. Thompson and Heath (1999) proposed a whole-meiosis Gibbs sampler, the M-sampler. Although new values for each inheritance indicator of $v_{\cdot,j}$ for a given meiosis are proposed in a sequential (backward) manner, they are updated jointly. Conversely to the L-sampler, the M-sampler mixes well with multiple linked loci, but mixes poorly on extended pedigrees. The combination of two samplers, LM-sampler, can improve the robustness and reliability of the MCMC estimates (Heath and Thompson 1997). At each step, either an M-sampler or an L-sampler can be randomly chosen based on a proportion specified in advance.

Application of likelihood on pedigrees in human genetics

One main application area of likelihoods on pedigree data is linkage analysis. As indicated in Chapter 1, a linkage analysis extracts cosegregation information of markers and disease from pedigree data. The majority of existing methods, either parametric or non-parametric, especially the former, is based on likelihood calculations. For parametric

linkage analysis with only a small number of loci, likelihood is most often calculated by the Elston-Stewart algorithm based on the assumed disease model. The parameter of interest is the recombination rate between markers and disease locus or location of the disease locus. When dealing with larger numbers of linked markers in parametric or non-parametric linkage analysis, people prefer the Lander-Green algorithm or Monte Carlo methods if pedigrees are large and complex.

Pedigrees are used for segregation analysis too. Segregation analysis tries to define the mode of inheritance of a Mendelian trait, and estimates the genetic model including the penetrance of each genotype, population allele frequencies, and transmission probabilities of a particular allele. All these can be estimated by maximizing over a likelihood. Likelihood ratio tests can be used to compare alternative models.

When traits of known mode of inheritance are observed, we can reconstruct genealogical structure based on the genetic data of related individuals. The observed joint phenotypes/genotypes among relatives provide information on their relationships since the joint probabilities of observed phenotypes/genotypes are functions of the underlying relationships among the subjects. The likelihood of a hypothetical relationship is the probability of the observed phenotypes/genotypes conditional on the known genetic model and assumed genealogical structure. By comparing the likelihood of alternative hypotheses, genealogical structure can be reconstructed as the most likely one. Genealogy reconstruction provides some practical applications. Sometimes, relationships between/among individuals are unreliable due to errors in collecting/recording data. The

ability to confirm the genealogy or detect errors is a key step in quality control before further analysis of the pedigree data. In linkage and association study for disease gene mapping, a pairwise relationship check is widely used due to its simplicity and efficiency. On the other hand, the accuracy of reconstructing an accurately known genealogy is a measure of the genetic diversity in a population. High similarities among different branches of the genealogy will cause errors when inferring true relationship based on genetic data.

Another area in human genetics using pedigrees is to using known pedigree structure and the mode of inheritance for a trait to compute the likelihood of unobserved phenotypes or genotypes given available observations. The unobserved phenotypes or genotypes may refer to potential offspring that are not born yet. This is a general question of interest in genetic counseling. Also we can estimate the genotypes or phenotypes of ancestors who are long since dead to specify allele origination. Inference on gene extinction and survival can come from ancestral likelihood of pedigrees.

References

Abecassis, G., Cherny, S., Cookson, W. and Cardon, L. (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30, 97-101

Baum, L. E. (1972) An inequality and associated maximization technique in statistical estimation for probabilistic function on Markov processes, *in* O. Shisha, ed., 'Inequalities-III; Proceedings to the Third symposium on Inequalities. University of California Los Angeles, 1969', Academic Press, New York, pp. 1-8

Cannings. C., Thompson, E. A. and Skolnick, M. H. (1978), Probability functions on complex pedigrees, *Adv. App. Prob.* 10, 26-61

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

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

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

Heath, S. C. and Thompson, E. A. (1997) MCMC samplers for multilocus analysis on complex pedigrees, *Am. J. Hum. Genet.* 61, A278

Heuch, I. and Li, F. M. H. (1972), PEDIG – A computer program for calculation of genotype probabilities, using phenotypic information, *Clinic. Genet.* 3, 501-504

Indury, R. and Elston, R. (1997) A faster and more general hidden Marko model algorithm for multipoint likelihood calculations. *Nat. Genet.* 47, 197-202

Hilden, J. (1970), GENEX – An algebraic approach to pedigree probability calculus, *Clinic. Genet.*1, 319-348

Iwin, M., Cox, N. and Kong, A. (1994) Sequential imputation for multilocus linkage analysis, *Proc. Nat. Acad. Sci.* (USA) 91, 11684-11688

Jensen, C. S., Kjaeruff, U. and Kong, A. (1995) Blocking Gibbs sampling in very large probabilistic expert systems, *Inter. J. Hum. –computer Studies* 42, 647-666

Kong, A. (1991) Analysis of pedigree data using methods combining peeling and Gibbs sampling, in E. M. Keramidas and S. M. Kaufman, eds, 'Computer Science and Statistics: Proceedings of the 23rd Symposium on the interface', Interface Foundation of North America, Fairfax Station, VA, pp. 379-385

Kong, A., Liu, J. and Wong, W. H. (1994) Sequential imputations and Bayesian missing data problems, *J. Am. Stat. Assoc.* 89, 278-288

Kruglyak, L., Daly, M. J. and Lander, E. S. (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping, *Am. J. Hum. Genet* 56, 519-527

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

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

Lander, E. S. and Green, P. (1987) Construction of multilocus genetic linkage maps in humans. *Proc. Nat. Acad. Sci.* (USA) 84, 2363-2367

Lathrop, G. M. , Hooper, A. B., Huntsman, J. W. and Ward, R. H. (1984) Strategies for multilocus linkage analysis in humans. *Proc. Nat. Acad. Sci.* (USA) 81, 3443-3446

Markianos, K., Daly, M. and Kruglyak, L (2001a) Efficient multipoint linkage analysis through reduction of inheritance space. *Am. J. Hum. Genet.* 68, 963-977

Markianos, K., Katz, A. and Kruglyak, L (2001b) a new computational approach for rapid multipoint linkage analysis of qualitative and quantitative traits in large, complex pedigrees, and its implementation in GENEHUNTER, *Am. J. Hum. Genet.* 69, 228

Metropolis, N., Rosenbluth, A. W., Roenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines, *J. Chem. Phys.* 21, 1087-1092

Skrivanek, Z., Lin, S. and Iwin, M (2003) Linkage analysis with sequential imputation *Genet. Epidem.* 25, 25-35

Thompson, E. A. (1986), *Pedigree Analysis in Human Genetics,* Johns Hopkins University Press, Baltimore

Thompson, E. A. (1994) Monte Carlo likelihood in genetic mapping, *Stat. Sci*. 9, 355-366

Thompson, E. A. (2000), Statistical Inference from Genetic Data on Pedigrees

Thompson, E. A. and Guo, S. W. (1991) Evaluation of likelihood ratios for complex genetic models, I.M.A. *J. Math. App. Med. Bio.* 8, 149-169

Thompson, E. A. and Heath, S. C. (1999) Estimation of conditional multilocus gene identity among relatives, in F. Sellier-Moiseiwitsch, ed., 'Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology', IMS Lecture Note-Monograph Series Volume 33, Institute of Mathematical Statistics, Hayward, CA, pp. 95-113

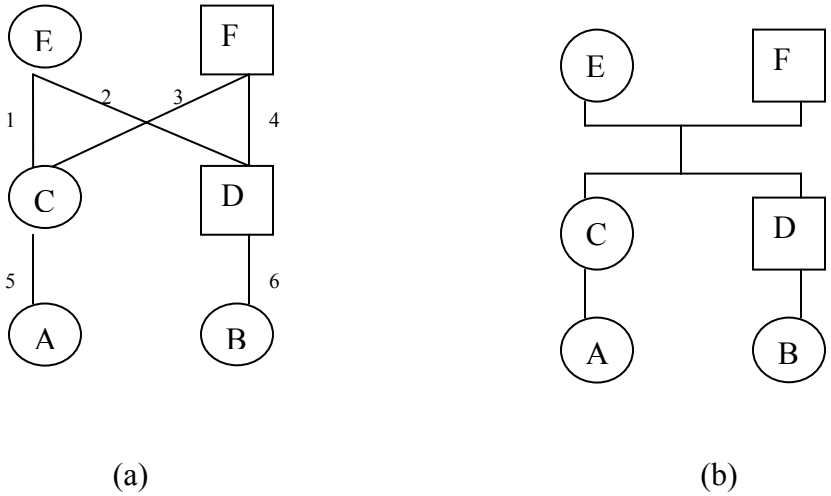


Figure 2.1, graphic presentations of a 3 generation pedigree.

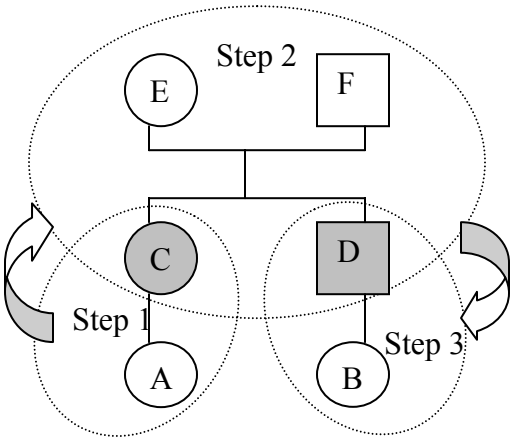


Figure 2.2. One possible peeling sequence of the "peeling" algorithm.

For the pedigree in figure 2.1b. Shaded individuals are pivots connecting a nuclear family to the rest of the pedigree. Arrows indicates moving directions along the pedigree.

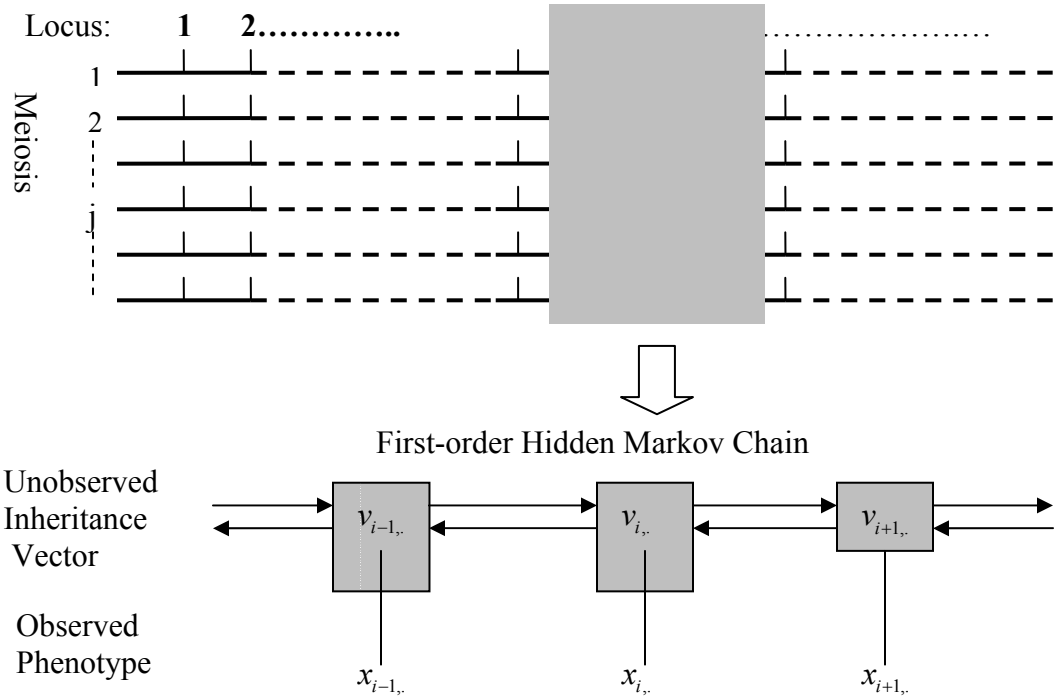


Figure 2.3 The first-order Hidden Markov Chain Structure for Lander-Green algorithm.

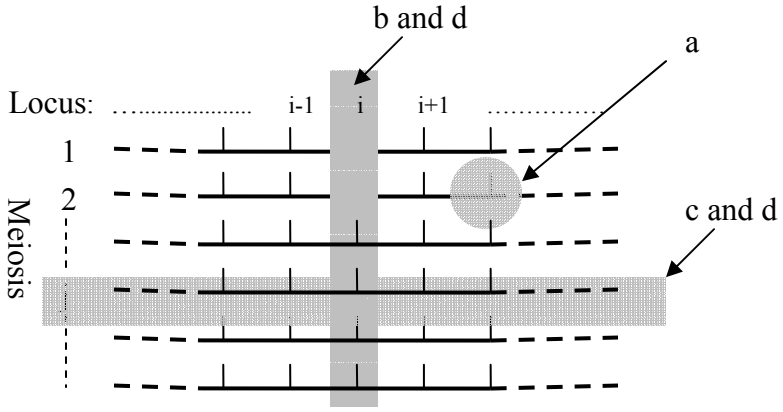


Figure 2.4 MCMC sampler using inheritance vectors as latent variable for general pedigrees.

Shaded area indicated a sample unit for different sampler.

a. Single-site sampler: updating an inheritance indicator
b. L-sampler: updating an inheritance vector $v_{i,.}$ at a locus together
c. M-sampler: updating inheritance vector $v_{.,j}$ for a meiosis together
d. ML-sampler: updating by combined L- and M-sampler.

Chapter 3

Sampling importance resampling estimation of likelihoods for pairwise relationship inference

Li Li

Bioinformatics Research Center

North Carolina State University

Raleigh, NC 27695-7566

and

Sharon Browning

GlaxoSmithKline

Research Triangle Park, NC 27709

Running heads: Relationship inference

Key words: relationship inference, sequential importance sampling, resampling

Address for correspondence:

Sharon Browning

GlaxoSmithKline

5 Moore Drive

Research Triangle Park, NC 27709

e-mail: Sharon.R.Browning@gsk.com

phone: 919-483-9010

Abstract

Traditional likelihood calculations for pedigree data analysis are based on algorithms that are infeasible when both pedigree size and number of markers are large. A Sequential Importance Sampling (SIS) approach has been proposed to allow the calculation for large pedigrees with large numbers of markers. However, when the system gets large, the variance of the importance sampling weights increases while efficiency and accuracy of the method decrease. In our calculation of likelihoods for pairwise relationship inference, we incorporate a resampling strategy into SIS to reduce the variance inflation problem. Instead of sampling each sampled path (inheritance pattern) along the chromosome independently, we sample a number of paths in parallel and after a certain number of marker loci, draw a new sample set from the current sample set according to their importance sampling weights. We compare the efficiency of our method to SIS on simulated data. Effective sample sizes are four to 100 times greater with resampling than with SIS only, with greatest gains for densely spaced markers.

Introduction

Probabilities of genotypic data from multiple linked marker loci on related individuals are used as likelihoods of gene locations for gene-mapping (Lander and Green, 1987; Irwin et al., 1994; Kruglyak and Lander, 1998; Thompson, 2001) or as likelihoods of relationship for relationship inference (Thompson, 1986; Boehnke and Cox, 1997; Epstein et al., 2000; Sieberts et al., 2002). These probabilities may be calculated exactly with the Elston-Stewart algorithm (Elston and Stewart, 1971) or Lander-Green algorithm (Lander and Green, 1987). Computational time for the Elston-Stewart algorithm scales linearly with the number of meioses in a pedigree and exponentially with the number of marker loci on a chromosome, while the Lander-Green algorithm scales linearly with the number of markers but exponentially with the number of loci.

In this paper, we concentrate on pairwise relationship inference, although the approach may be generalized to inference of relationships between multiple individuals or to linkage gene-mapping. Pairwise relationships between individuals play an important role in linkage and population studies. In linkage studies, pairwise relationships may be examined to check for sample swaps or mis-reported relationships (Boehnke and Cox, 1997; Epstein et al., 2000). In population studies, pairwise relationships may be used to reconstruct genealogies, which are useful for studying population structure or for further genetic investigation (Thompson, 1975). Hence we need accurate and efficient methods for pairwise relationship inference.

A likelihood ratio testing approach may be applied to pairwise relationship inference. Data consist of genetic marker genotypes at multiple linked loci. If we also knew the underlying inheritance patterns, i.e., exactly which allele in an individual's genotype is inherited from the mother and which from the father, the likelihood calculation would be simple. However the inheritance patterns are generally unobserved. Exact methods for pedigree data analysis need to sum out all combinations of inheritance patterns over all loci. For m meioses in a pedigree and l markers, there are 2^{m*l} combinations so a direct approach to the summation would be infeasible when m or l is large.

In the context of linkage analysis, Lander and Green (1987) showed that the computational order of the sum can be reduced to $l4^m$, while refinements to their algorithm have further reduced the computational order to $lm2^m$ (Kruglyak and Lander, 1998). Boehnke and Cox (1997) used a similar method to calculate likelihoods of first-degree pairwise relationships (full-sibs, half-sibs, monozygotic twins and unrelated pairs). Epstein et al. (2000) extended the method to test second-degree relationships: grandparent-grandchild, half-sib and avuncular pairs, using approximate likelihoods for avuncular pairs.

Many terms in the exact likelihood sum are extremely small, as the corresponding inheritance patterns are shown to be very unlikely given the observed genotype data. An importance sampling approach aims to obtain a good estimate of the likelihood by sampling terms in the sum, concentrating on those terms with largest values in order to minimize the variability of the estimate. Irwin et al. (1994) proposed a Monte Carlo

sequential importance sampling (SIS) approach to likelihood calculation for linkage analysis with large pedigrees with a large number of marker loci. The algorithm involves sampling of inheritance indicators for all meioses at all loci, proceeding sequentially along the chromosomes. However, for a fixed number of samples, the variance of the estimated likelihoods increases and the accuracy of inference decreases when pedigree size and number of loci increase.

In this paper, we consider a resampling strategy to improve the performance of SIS in calculating likelihoods of pairwise relationships. Rather than sampling every realization of the inheritance indicators sequentially and independently along the chromosome via SIS, we sample N realizations simultaneously along the chromosome and at certain intervals draw a new set of samples based on weights of the realizations. We present the results of a simulation study in which we compare SIS and our resampling approach. Comparing the results, for a given number of samples our method decreases the variation coefficient of the estimate by up to an order of magnitude, depending on marker interval spacing. The greatest improvements are seen with densely spaced markers.

Methods

Assumptions and definitions

We assume Hardy-Weinberg equilibrium and no selection or mutation. Let l be the number of loci, $X = (x_1, \ldots, x_l)$ denotes the observed genotypes at the marker loci for the pair of individuals we are interested in, where x_i represents the genotypes of the pair at locus i. We assume that the genotypes are recorded without error. Let r_i be the recombination rate between markers i and $i+1$. We assume r_i is known and is the same for both sexes and that marker order is known. Only autosomal chromosomes are considered here. Allele frequencies are also assumed known or can be estimated from data. We assume no crossover interference or linkage disequilibrium. Given a hypothesized relationship θ, let m be the number of meiosis in the pedigree describing this relationship. For a pair of individuals, we will consider several possible relationships. For example, in our simulation study we consider the relationships MZ twins, full siblings, parent-offspring, unrelated, grandparent-grandchild, aunt-niece, half siblings and first cousins. Each of these relationships in turn is considered as the hypothesized relationship θ, and the likelihood of this relationship is calculated (or estimated). The relationship with the highest likelihood is inferred to be the true relationship.

Inheritance vectors and the likelihood

Given a hypothesized relationship θ corresponding to a pedigree with m meioses, an inheritance vector v_i is a binary vector of length m representing inheritance information at locus i. The j th coordinate of the inheritance vector corresponds to the j th

meiosis, $j = 1,\ldots m$. Each meiosis involves a parent and child: an inheritance value of 0 for the meiosis indicates that the child inherited the parent's maternal allele (the allele he or she had inherited from his or her mother) at the locus, while 1 indicates the allele was inherited from parent's paternal allele. These vectors are "missing data" as they are generally unobservable, and each has $s = 2^m$ possible values. For a hypothesized relationship θ, the true inheritance vector for each marker locus has the same length and structure but may be of different value due to recombination between loci.

Given the relationship between the individuals, an inheritance vector determines the identity by descent (IBD) status of the individuals at the locus. For example, consider the avuncular relationship in Figure 3.1a, with 5 meioses. The inheritance vector is of length 5 and has $s = 32$ possible values. One possibility is (0, 0, 1, 1, 0) which means individual A inherits her mother (C)'s maternal allele in meiosis 1, and her father (E)'s paternal allele in meiosis 3, while individual D inherits his mother (C)'s maternal allele in meiosis 2, and his father (E)'s paternal allele in meiosis 4, and individual B inherits her father (D)'s maternal allele in meiosis 5, which comes from C's maternal allele. Thus A's two alleles are C's maternal and E's paternal allele, while B's two alleles are C's maternal allele (from D's maternal allele) and an allele from B's mother who is unrelated to the other individuals in the pedigree. Thus A and B share one allele (C's maternal allele) identical by descent (IBD). For the grandparent-grandchild relationship in Figure 3.1b, the inheritance vector is of length two and has four possible values. If the value of the inheritance vector is (0, 0), it means individual C inherits her mother (A)'s maternal allele in meiosis 1, while individual B inherits his mother (C)'s maternal allele in meiosis 2,

which comes from A's maternal allele. Thus A and B share one allele (A's maternal allele) IBD.

Figure 3.1. a and b will be placed here.

Under the assumption of no crossover interference, the inheritance vectors form a first order Markov Chain moving along the loci. Given the hypothesized relationship θ, let $V = (v_1, \ldots, v_l)$ be the underlying inheritance patterns at the marker loci, where hidden state v_i is the inheritance vector for locus i, $i = 1, \ldots, l$. We name each possible inheritance pattern V "an inheritance path". Define $|v_j - v_k|$ to be the number of coordinates that differ between v_j and v_k. Each coordinate that differs between v_j and v_k represents a recombination in the corresponding meiosis between marker loci j and k. Thus the transition probability $P_\theta(v_{i+1} \mid v_i)$ is $r_i^{|v_{i+1} - v_i|}(1 - r_i)^{m - |v_{i+1} - v_i|}$. The a priori initial probability for each vector $P_\theta(v_1)$ is $1/s$, because each inheritance vector is equally probable a priori. Given allele frequencies and the assumption of Hardy-Weinberg equilibrium, the probability of observed genotypes x_i at locus i given inheritance vector v_i $P_\theta(x_i \mid v_i)$ can be calculated according to the corresponding IBD status of v_i (Thompson, 1975; Epstein et al., 2000). These emission probabilities are given in Table 3.1, and the connection between IBD status and inheritance vector for the avuncular and grandparent-grandchild relationships from Figure 3.1 is given in Table 3.2.

Table 3.1 and Table 3.2 will be placed here.

The full probabilistic model we consider describes the probability of the observed genotype data X given the relationship θ. This probability may also be thought of as the likelihood of relationship θ given genotype data X. Our goal is to find the relationship that maximizes this likelihood. From the discussion above, we see that to calculate the probability of the genotype data (or likelihood of the relationship), we need to sum over possible values of the "missing data", the unobserved inheritance path V:

$$\begin{aligned} L(\theta \mid X) &= P_\theta(X) \\ &= \sum\nolimits_V P_\theta(X, V) \\ &= \sum\nolimits_V P_\theta(X \mid V) P_\theta(V) \\ &= \sum\nolimits_V \left[\prod_{i=1}^{l} P_\theta(x_i \mid v_i) \right] \left[P_\theta(v_1) \prod_{i=1}^{l-1} P_\theta(v_{i+1} \mid v_i) \right] \end{aligned}$$

The components $P_\theta(x_i \mid v_i)$, $P_\theta(v_1)$ and $P_\theta(v_{i+1} \mid v_i)$ are described above.

Sequential Importance Sampling

Importance sampling is a Monte Carlo sampling approach which focuses on "importance" regions to save computational resources. Sequential importance sampling (SIS) is a special case of importance sampling that is most useful in state-space models. It builds up the sampling distribution sequentially and draws samples recursively. Ideally, we want to draw samples from the conditional distribution $P_\theta(V \mid X)$. But $P_\theta(V \mid X)$ is usually as hard to calculate as or even more difficult than the likelihood itself. However, samples can be drawn from some easy-to-sample trial distribution P_θ^* with a similar shape to the conditional distribution $P_\theta(V \mid X)$ in order to make the estimation with a small standard error, provided that proper weights are assigned to samples (Thompson, 2001; Liu, 2001).

Let X_i, V_i denote the observed data and inheritance pattern up to locus i respectively. Here, we impute v_i sequentially from a trial distribution P^* close to the conditional distribution $P_\theta(V \mid X)$, and then correct the bias by using importance weights to estimate likelihood. Two steps are involved:

(1) Draw v_1^* from $P_\theta(v_1 \mid X_1)$, and then for $i = 2,\ldots,l$, draw v_i^* sequentially from the conditional distribution $P_\theta(v_i \mid X_i, V_{i-1}^*)$.

(2) Sequentially compute $P_\theta(x_i \mid X_{i-1}, V_{i-1}^*)$ and importance weight $w_i = w_{i-1} P_\theta(x_i \mid X_{i-1}, V_{i-1}^*)$, where $w_1 = P_\theta(X_1) = P_\theta(x_1)$.

These two steps are performed simultaneously for each locus. The probabilities $P_\theta(v_i \mid X_i, V_{i-1}^*)$ and $P_\theta(x_i \mid X_{i-1}, V_{i-1}^*)$ can be calculated sequentially using initial probabilities, emission probabilities and transition probabilities.

$$P_\theta(X_1) = \sum_{v_1} P_\theta(x_1 \mid v_1) P_\theta(v_1)$$
$$\text{and } P_\theta(x_i \mid X_{i-1}, V_{i-1}^*) = \sum_{v_i} P_\theta(x_i \mid v_i) P_\theta(v_i \mid v_{i-1}^*) \text{ for } i = 2,\ldots,l$$

$$P_\theta(v_1 \mid X_1) = \frac{P_\theta(x_1 \mid v_1) P_\theta(v_1)}{P_\theta(X_1)}$$
$$\text{and } P_\theta(v_i \mid X_i, V_{i-1}^*) = \frac{P_\theta(x_i \mid v_i) P_\theta(v_i \mid v_{i-1}^*)}{P_\theta(x_i \mid X_{i-1}, V_{i-1}^*)} \text{ for } i = 2,\ldots,l$$

Let $W = w_l$ and $V^* = (v_1^*,\ldots,v_l^*)$ be the weight and sampled inheritance path of a single SIS run. Suppose N paths are sampled independently, the results will be

$\{V^*(1),\ldots,V^*(N)\}$ and $\{W(1),\ldots,W(N)\}$. It can be shown that an unbiased estimator of $L(\theta \mid X) = P_\theta(X)$ is $\overline{W} = \frac{1}{N}\sum_{j=1}^{N} W(j)$ (Irwin et al., 1994; Liu, 2001).

Sampling-Importance Resampling (SIR)

The basic idea of importance sampling suggests that we should pay more attention to the regions of importance (Marshall, 1956). However after a number of steps, or more and more data being processed, the importance sampling weights of some sampled paths can become too small, increasing the estimation variance. Those paths with relatively small weights are said to be "ineffective" because they contribute little to the final estimation. The Monte Carlo computation will be less efficient as the number of ineffective sample paths increases (Liu and Chen, 1995; Liu, 2001). To deal with this difficulty, we propose a resampling strategy to discard the ineffective samples partway through the sequential sampling process (as in Liu, 2001). Suppose we do resampling at locus i:

(1) Independently sample N paths simultaneously by SIS up to locus i. The current results will be $\{V_i^*(1), V_i^*(2),\ldots,V_i^*(N)\}$ and $\{w_i(1), w_i(2),\ldots,w_i(N)\}$, where $V_i^*(j)$, $w_i(j)$ represent the partial sampled path and weight up to locus i for the jth path, $j = 1,2,\ldots,N$.

(2) Resample a new set of partial sampled paths from $\{V_i^*(1), V_i^*(2),\ldots,V_i^*(N)\}$ according to the current weights $\{w_i(1), w_i(2),\ldots,w_i(N)\}$. The N original partial sampled paths will be replaced by the new set denoted by $\{V_i'(1), V_i'(2),\ldots,V_i'(N)\}$

with equal weights $w_i^{'}(j) = \frac{1}{N}\sum_{k=1}^{N} w_i(k)$, $j = 1, 2, \ldots, N$. This resampling step can be done via two different resampling strategies:

a) Simple Random Resampling(SRR)

Sample from $\{V_i^*(1), V_i^*(2), \ldots, V_i^*(N)\}$ with probabilities proportional to their weights $\{w_i(1), w_i(2), \ldots, w_i(N)\}$

b) Residual Resampling(RR)

Keep $c_j = \left[\frac{N w_i(j)}{\sum_{k=1}^{N} w_i(k)} \right]$ (where [] represents the floor function) copies of $V_i^*(j)$, $j = 1, 2, \ldots, N$, and further obtain $c_r = N - \sum_{j=1}^{N} c_j$ independent samples from $\{V_i^*(1), V_i^*(2), \ldots, V_i^*(N)\}$ with probabilities proportional to $\frac{N w_i(j)}{\sum_{k=1}^{N} w_i(k)} - c_j$, $j = 1, 2, \ldots, N$.

(3) Continue the SIS for the remaining loci based on the resampled *N* paths and new weights.

The paths with large weights are more likely to be resampled and kept, while those ineffective ones will be discarded because of low resampling probabilities. Resampling can be done at every step of SIS. But resampling too often will increase the computational burden. We can choose a resampling schedule according to the situation. When the calculation of the variance coefficient of weights (cv^2) is computationally

efficient, a dynamic resampling step can be done when the effective sample size (see below) is less than some threshold value. Or if the calculation is time-consuming, we can do resampling at predetermined stages, such as every five or 10 loci (Liu, 2001).

In the resampling step, new samples are drawn from the current samples. No previously unsampled paths can be drawn from the trial distribution. As the sampling proceeds along the chromosome, at some stage the estimate from the samples may no longer adequately represent the likelihood (Berzuini et al., 1997). Furthermore, the samples become dependent after resampling steps (Doucet and Andrieu, 2000). We can solve these problems by running the data in batches. Given the total number of simulations, we can divide them into b batches, processing them independently. A mean weight can be obtained for each single batch using the approach described above. The final estimate is just the average of these batch means.

Effective Sample Size

Effective sample size (ESS) was defined by Kong et al. (1994) for sequential importance sampling to measure how different the trial distribution is from the target distribution. The more similar the trial and target distributions are, the larger the ESS is and the more efficient is the sequential importance sampler.

$$ESS(N) = \frac{N}{1 + Var_{P_\theta^*}(W^*)}$$

where W^* is the weight normalized by $\overline{W} = \frac{1}{N}\sum_{j=1}^{N} W(j)$, and has $EW^* = 1$. Thus the coefficient of variation of the unnormalized weight, $cv^2(W)$, estimates the variance of

W^*. On the other hand, $cv^2(\overline{W}) = \frac{1}{N} cv^2(W)$ can be approximated by the standard error squared of the natural log of the average weight $SE^2(\ln \overline{W})$ based on the delta method which approximates the variance of $\ln \overline{W}$ by first expanding the function $\ln \overline{W}$ with a 1-step Taylor approximation, and then taking the variance such that

$$SE^2(\ln \overline{W}) \approx \frac{Var(\overline{W})}{\overline{W}^2} = cv^2(\overline{W}).$$

Hence, when N is large,

$$\widehat{ESS}(N) \approx \frac{1}{SE^2(\ln \overline{W})}$$

which means $SE(\ln \overline{W})$ measures the efficiency of the sampler, where a smaller value corresponds to higher efficiency. Since all calculations and values in the computer implementation are on natural log scale to avoid underflow, using $SE^2(\ln \overline{W})$ to evaluate the efficiency of the sampler simplifies the calculation compared to using $cv^2(W)$.

Simulations

We generated 50 pairs of avuncular individuals for marker loci equally spaced along the 22 autosomal chromosomes (genetic length information was obtained from Marshfield comprehensive human genetic maps) at 5 cM, 10 cM, 20 cM or 35 cM intervals. Markers are microsatellites each with seven alleles with frequencies: .40, .20, .20, .05, .05, .05, and .05. The estimated likelihoods for the SIR and SIS methods were in each case based on 10^5 samples divided into 1000 batches.

For each likelihood estimation method and pair of individuals we estimated the likelihoods of each of eight relationships: MZ twins (Figure 3.2a), full siblings (Figure 3.2b), parent-offspring (Figure 3.2c), unrelated (Figure 3.2d), half siblings (Figure 3.2e), first cousins (Figure 3.2f), grandparent-grandchild (Figure 3.1b) and avuncular (Figure 3.1a). We consider only these first and second degree relationships to make calculation of the exact likelihoods (for comparison with the estimated likelihoods) feasible. For a given estimation method, the inferred relationship is the one with the highest likelihood. Note that even with exact likelihoods, the relationship with the highest likelihood is not necessarily the true relationship, but, in the absence of prior information, picking the relationship with the highest likelihood is the best that we can do.

A confidence interval for an estimated likelihood reflects our uncertainty about the exact value of the likelihood. As the number of sampling replications increases, the confidence interval narrows. If the confidence intervals for the relationships with the highest and second highest estimated likelihoods do not overlap, we can be fairly confident that the

relationship with the highest estimated likelihood is also the relationship with the highest actual likelihood. However when these two confidence intervals overlap, we cannot be sure that we have determined which relationship has the highest actual likelihood, and we say that the results are ambiguous. One measure for comparing likelihood estimation schemes is the proportion of ambiguous results they produce.

Results

We used the computer simulation study described above to test the efficiency of our method (SIR) and compare it with the SIS method and to compare the estimated likelihoods with the actual likelihood**.**

First we applied deterministic resampling schedules at every 1, 2, 5, 10, or 15 loci for both Simple Random Resampling (SRR) and Residual Resampling (RR). Note that for some combinations of marker spacing and resampling schedule, some of the chromosomes didn't have any resampling. Regardless of the marker interval spacing, in most cases resampling every five loci for both SRR and RR gave the smallest standard error of $\ln \overline{W}$. Occasionally resampling every 10 or two loci gave the smallest standard error of $\ln \overline{W}$ but the value was very close to resampling every five loci. The results indicate that after a small number of sampling steps, when importance sampling weights are nearly the same for all samples, resampling reduces only the number of possible paths and introduces extra variation. On the other hand, after a large number of sampling steps, importance sampling weights become very diverse, so that only a small number of paths make it through the resampling step. In this case samples in a batch focus on only part of the "importance" region and the variance of final estimate increases. For the following deterministic resampling simulation study, resampling every five loci has been used.

Second, we applied dynamic resampling when effective sample sizes were less than 0.5, 0.75 or 0.9 of the original sample size. The average numbers of loci between resampling

points are comparable to different deterministic resampling schedules: 0.5 threshold is equivalent to resampling at every 10 to 15 loci, 0.75 threshold to resampling at every five loci and 0.90 threshold to resampling at every 2 loci. Since calculation of ESS at each locus is very time consuming, we chose to use only deterministic resampling for the remainder of our simulation study.

To test the efficiency and accuracy of SIR, we resampled every 5 loci using SRR and RR and compared the estimated 95% confidence intervals for the likelihoods under this method and under SIS to the exact likelihoods. For each hypothesized relationship, all confidence intervals from SRR (SIR), RR (SIR) and SIS contained the value of the exact likelihood. The standard error $SE(\ln \overline{W})$ for the SIR estimates was about a half to a tenth of that with SIS, which implies the effective sample size with SIR is about 4 to 100 times that with SIS. Although the absolute value of $SE(\ln \overline{W})$ increases as the marker density increases, the relative decrease in $SE(\ln \overline{W})$ from the SIR method compared to that from SIS is proportional to the density of markers. In other words, SIR is more efficient for data with closer marker interval spacing.

Table 3.3 shows the relationship inference results for the first 15 simulated avuncular pairs for marker spacing of 5cM. In cases where the results are unambiguous (confidence intervals for the likelihoods of the two relationships with the highest estimated likelihoods do not overlap), the inferred relationship (relationship with the highest exact or estimated likelihood) is the same with the estimation methods as with the exact

likelihoods, although the inferred relationship is not necessarily the true relationship due to insufficient information in the genotypic data.

Table 3.3 will be placed here.

We found that ambiguous results due to overlapping confidence intervals can be a problem with this number of sampling iterations, especially for second-degree relationships (grandparent-grandchild, avuncular, half siblings). As we can see in Table 3.4, the number of ambiguous results due to overlapping confidence intervals is less with SIR than with SIS. The degree of improvement is also proportional to the density of markers. For 5 cM interval space data, nearly no overlapping confidence intervals resulted from SIR while more than a half of the 50 simulated data sets gave overlapping intervals for the most likely relationship with SIS. This implies that SIR can give more confident conclusions than SIS, especially with dense markers.

Table 3.4 will be placed here.

Discussion

We incorporated a resampling strategy in sequential importance sampling in the context of estimating likelihoods of relationship from genotypic data. Resampling at about every five loci has been shown to be most efficient. The simulation study to compare results of SIS and our SIR method demonstrated that the performance of SIS is greatly improved by incorporating the resampling strategy, especially for data with dense markers. Both efficiency, which is measured by ESS, and accuracy, which is evaluated by the number of overlapping confidence intervals for the highest likelihoods, increase significantly. These results suggest that SIR will be much more efficient and accurate than SIS for analysis of large pedigrees with a large number of loci.

For simplicity in this paper we made a number of assumptions including Hardy-Weinberg equilibrium, no genotyping error and no sex differences in recombination rates. Further work could allow for genotyping errors, sex differences in recombination rates and X-linked marker data and Hardy-Weinberg disequilibrium. Also relationships can be inferred more accurately by considering multiple individuals together (Sieberts et al., 2002). The dimension of the inheritance vector space will tend to be higher for multiple individuals than for pairs of individuals due to large pedigrees. Thus, SIR would be appropriate since it is most useful for high-dimensional calculations in which other likelihood calculation methods are infeasible.

Furthermore, instead of inferring relationships by assuming known parameters, such as recombination frequencies between markers, we could use SIR to estimate these

parameters when the pedigree structure is known, so that this approach could be used for linkage and disease gene mapping. Skrivanek et al (2003) apply SIS to non-parametric linkage analysis. Their method works well for large pedigrees but is limited in the number of marker loci that can be considered. Combining their approach with our resampling would allow for calculations on increased numbers of loci.

Acknowledgments

Li Li is grateful for Kejun Liu for many helpful discussions in computational programming.

References

Berzuini, C., Best, N. G., Gills, W. R. and Larizza, C. 1997 Dynamic conditional independence models and Markov Chain Monte Carlo methods, *Journal of the American Statistical Association* 92: 1403-1412.

Boehnke, M and Cox, N. J. 1997 Accurate inference of relationships in sib-pair linkage studies, *American Journal of Human Genetics* 61: 423-429.

Doucet, A., S. J. and Andrieu, C. 2000 On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing* 10: 197-208.

Elston, R. C. and Stewart, J. 1971 A general model for the genetic analysis of pedigree data, *Human Heredity* 21: 523-542.

Epstein, M. P., Duren, W. L. and Boehnke, M. 2000 Improved inference of relationship for pairs of individuals, *American Journal of Human Genetics* 67: 1219-1231.

Irwin, M., Cox, N. and Kong, A. 1994 Sequential imputation for multiple linkage analysis, *Proceeding of National Academy of Sciences (USA)* 91: 11684-11688.

Kong, A., Liu, J. S, and Wong, W. H. 1994 Sequential imputation and Bayesian missing data problems, *Journal of the American Statistical Association* 89: 278-288.

Kruglyak, L., and Lander, E. S. 1998 Faster Multipoint Linkage Analysis Using Fourier Transforms, *Journal of Computational Biology* 5:1-7.

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

Liu, J. S. 2001 *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York.

Liu, J. S. and Chen, R. 1995 Blind deconvolution via sequential imputations, *Journal of the American Statistical Association* 90: 567-576.

Marshall, A. W. 1956 "*The use of multi-stage sampling schemes in Monte Carlo computations*" in Symposium on Monte Carlo Methods. New York, John Wiley. pp. 123--140.

Sieberts, S. K., Wijsman, E. M. and Thompson, E. A. 2002 Relationship inference from trios of individuals, in the presence of typing error*, American Journal of Human Genetics* 70: 170-180.

Skrivanek, Z., Lin, S. and Irwin, M. 2003 Linkage analysis with sequential imputation, *Genetic Epidemiology* 25:25-35.

Thompson, E. A. 1975 The estimation of pairwise relationships, *American Journal of Human Genetics* 39: 173-188.

Thompson, E. A. 1986 *Pedigree Analysis in Human Genetics*, Johns Hopkins University Press, Baltimore, Maryland.

Thompson, E. A. 2001 *Statistical Inferences from Genetic Data on Pedigrees*, Institute of Mathematical Statistics, Beachwood, Ohio.

Table 3.1 Emission probabilities $P_\theta(x_i \mid v_i)$ for ordered genotype pairs.

Genotype x_i **	IBD status * of inheritance vector v_i		
	0	**1**	**2**
$(a_i a_i,\ a_i a_i)$	p_i^4	p_i^3	p_i^2
$(a_i a_i,\ a_i a_j)$	$2p_i^3 p_j$	$p_i^2 p_j$	0
$(a_i a_i,\ a_j a_j)$	$p_i^2 p_j^2$	0	0
$(a_i a_i,\ a_j a_k)$	$2p_i^2 p_j p_k$	0	0
$(a_i a_j,\ a_i a_j)$	$4p_i^2 p_j^2$	$p_i p_j (p_i + p_j)$	$2p_i p_j$
$(a_i a_j,\ a_i a_k)$	$4p_i^2 p_j p_k$	$p_i p_j p_k$	0
$(a_i a_j,\ a_k a_h)$	$4p_i p_j p_k p_h$	0	0

*IBD status represents the numbers of alleles that are identical by descent between two non-inbred individuals

** $a_i, a_j,\ a_k, a_h$ are distinct alleles with allele frequencies p_i, p_j, p_k, p_h respectively.

Table 3.2. Corresponding IBD status for each inheritance vector.

IBD status	relationship	0	1	2
Inheritance Vectors	aunt-niece	All other possible vectors	(0,0,a*,b*,0) (1,1,a,b,0) (a,b,1,1,1) (a,b,0,0,1)	NA
	Grandma-granddaughter	(a, 1)	(a, 0)	NA

* a and b indicate either 0 or 1.

Table 3.3. Relationship inference for first 15 simulated avuncular pairs.

Relationship	Likelihood estimation method			
	Exact likelihood	SIS	SRR	RR
1	Half siblings	?(Half siblings/Grandpa-grandchild)	Half siblings	Half siblings
2	Avuncular	?(Avuncular/Half siblings)	Avuncular	Avuncular
3	Avuncular	Avuncular	Avuncular	Avuncular
4	Avuncular	Avuncular	Avuncular	Avuncular
5	Half siblings	Half siblings	Half siblings	Half siblings
6	Avuncular	?(Avuncular/Half siblings)	Avuncular	Avuncular
7	Avuncular	?(Avuncular/Half siblings)	Avuncular	Avuncular
8	Avuncular	?(Avuncular/Half siblings)	Avuncular	Avuncular
9	Avuncular	Avuncular	Avuncular	Avuncular
10	Avuncular	?(Avuncular/Half siblings)	Avuncular	Avuncular
11	Half siblings	?(Avuncular/Half siblings)	Half siblings	Half siblings
12	Half siblings	?(Avuncular/Half siblings)	Half siblings	Half siblings
13	Avuncular	?(Avuncular/Half siblings)	Avuncular	Avuncular
14	Avuncular	?(Avuncular/Half siblings)	Avuncular	Avuncular
15	Avuncular	Avuncular	Avuncular	Avuncular

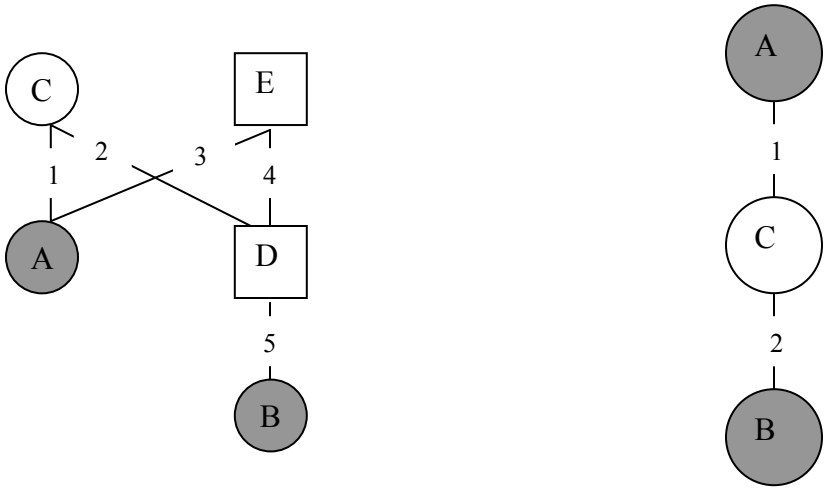
At 5cM marker spacing

Table 3.4. Number of ambiguous results out of 50 simulated avuncular pairs.

Interval spacing		35cM	20 CM	10 cM	5 cM
Number of ambiguous results	SIS	3	8	17	30
	SRR	3	3	3	1
	RR	2	3	2	0

Ambiguous: overlapped confidence intervals for highest likelihoods

.



a b

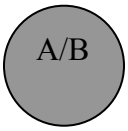
Figure 3.1 Examples of relationships between individuals A and B.

Individuals are shown by circles (females) and squares (male). The shaded circles are the individuals of interest with known genotypes. All other individuals are unknown. Meioses are shown by lines and labeled with integers.

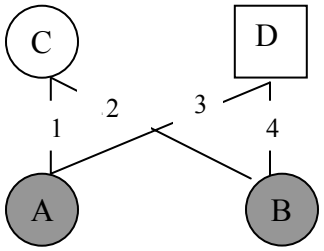
a. Avuncular (aunt-niece) relationship

b. Grandparent-grandchild (Grandma-granddaughter) relationship

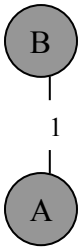
a. MZ twins



b. full siblings

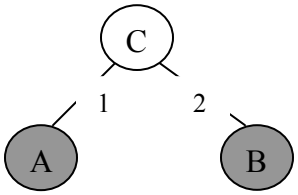


c. parent-offspring



d. unrelated

e. half siblings



f. first cousins

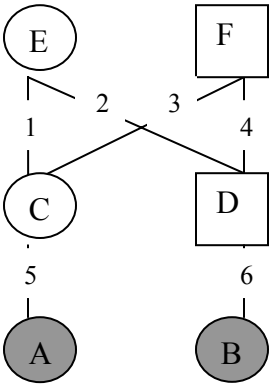


Figure 3.2 More examples of relationships between two individuals.

(Notation is the same as in Figure 3.1)

Chapter 4

Association explaining non-parametric linkage for binary traits in general pedigrees

Abstract

A successful linkage analysis may identify a linkage region spanning 10-30 MB or more containing hundreds of genes at a magnitude. Identifying the responsible gene can be accomplished using an association study, although a SNP or SNPs with an overwhelmingly small p-values rarely emerge. When several weak association signals are found within a linkage region, one of the key questions is whether the false positives can be distinguished from true association signals. We present a method for determining whether an associated SNP is responsible for a non-parametric linkage result for a binary trait in general pedigrees. The correlation between family frequency of a variant of interest and family LOD score is used as a measure of whether the association between a given variant at a marker and the disease status can help to explain a significant linkage result seen in the collection of families in the region around the marker. Results from simulation studies indicate that the method we proposed has a valid type I error rate for most of the tested disease models. We studied the effects of allele frequencies at a disease locus, allele frequencies at tested marker, and LD between a tested locus and a disease locus, different test types (allelic, genotypic and combined tests). With fixed population prevalence and phenocopy rate for a disease, increasing disease-causing allele frequency results in a reduced power of our method as it reduces the relative risk for disease-causing allele/genotype(s). Different test types fit different disease models well. The genotypic test is superior for the additive and dominant disease models, while the allelic model is slightly better than the genotypic test for recessive models. But the combined test has a performance very close to the preferred one in general. While we describe

results for a single marker test for an autosomal region, it is possible to extend the test to a haplotype-based one.

Introduction

Non-parametric linkage analysis methods do not assume a disease model. These methods test the hypothesis of excess identity by descent (IBD) sharing among affected individuals in each pedigree. The IBD sharing is not allele specific. Pedigrees segregating different alleles may contribute to the same linkage signal. A linkage region of interest will often contain hundreds of genes with 10-30 MB or more.

Association (or so-called linkage disequilibrium) methods test the hypothesis that an allele or genotype is correlated with affection status. Population based association methods examine the difference of frequencies of a variant between cases and controls, while family based methods try to identify over-transmission of a variant of interest by using the untransmitted one as controls. Positive association signals usually extend over regions less than 1 MB.

Association study results often identifying signals with modest p-values. Identifying the true signals from false positive results is not straight forward. Testing whether an associate variant (allele/genotype/haplotype) of interest tends to follow the IBD sharing pattern and the affection status in all pedigrees can provide additional information. We adopt the term "Association explaining linkage" for this new approach that ties linkage and association methods together. If a single variant is the only functional variant for a trait, the affected individuals in all families would show excess IBD sharing for the variant. We say "Association fully explaining linkage" for the disease variant in this situation. However, available markers may not be the true disease variant but be in LD

(not complete LD) with the disease variant for most of the time. Further, it is common that multiple loci with small marginal effects are responsible for a common disease. Such an individual variant can not fully explain an observed linkage signal. Methods that solve the problem of "Association partially explaining linkage" are very useful in the context of complex traits.

"Association explaining linkage" methods have some requirements for markers and data. Pedigrees are required as only such data provide information on linkage. Besides markers genotyped in pedigrees for linkage analysis, additional markers for association tests need to be genotyped in the same family samples in the linkage region of interest.

Sib-pairs are used by pioneers in this area. Horikawa et al (2000) and Cox (2001) consider the partial explanation problem for non-parametric qualitative linkage results using sib-pairs. One of the approaches they present is looking for correlation between the genotype in the patient, who is selected from each sib-pair first, and the LOD score for the sib-pair. It can be easily generalized to other pedigree structures, however it does not make use of full genotype information that may be available for the pedigree. The selection scheme of one case from each family should be carefully considered for general pedigrees (Li et al, 2004), although this is not an issue for sib-pairs. In the second approach, they identify a subset of sib-pairs where both sibs are concordant for a genotype of interest and compare the LOD score for the subgroup with the overall LOD score from all sib-pairs. This approach has been generalized to larger sibships but has difficulties to extend to other more complex pedigrees.

In the context of "Association fully explaining linkage", Sun et al (2002) address the problem using non-parametric linkage analysis for binary traits. They test the hypothesis whether a tested locus is the single site responsible for the trait or not by examining the conditional distribution of IBD status of an ASP given their genotypes at the tested locus is the same as that of a random sib-pair. Again this method uses sib-pairs and does not easily generalize to other pedigree structures. Besides non-parametric linkage, other authors work on using association to explain parametric linkage results. Li and Boehnke (2003) fit a parameter for the disequilibrium between the tested locus and the disease-susceptibility locus and test whether the disequilibrium is zero or not to see whether the association does not help to explain the linkage result or the association fully explains the linkage result. Parametric linkage analysis requires a model of inheritance of a disease. For many complex traits, the underlying disease model is normally unknown and hard to model, likely affecting the performance of their methods.

Cardon and Abecasis (2000) and several other authors have looked at the full explanation question in the context of quantitative traits. A variance components model can be fitted with parameters for linkage and an additional parameter for the locus of interest, then the hypothesis of whether a variant is the causal variant or is merely in LD with the causal variant can be tested by examining whether any linkage signal remains after accounting for the association.

Our approach assumes general pedigrees and tests for correlation between the frequency of a variant in affected individuals in a pedigree and the nonparametric LOD score for that family. This addresses the problem of whether an association partially explains the linkage result. In this paper, we focus primarily on allele and genotype frequencies. Allelic, genotypic and combined tests are proposed and compared. We will outline how to extend the test to a haplotype-based one. We also estimate the type I error rate under different disease models and study the effects of allele frequencies at a disease locus, allele frequencies at tested marker, and LD between a tested locus and a disease locus.

Methods

A linkage or LOD score peak is identified as a position/region with the maximum LOD score summed over all families. It is not necessary to have a genotyped marker at the peak. The individual LOD score for each family is recorded at the linkage peak. The tested marker is defined as the variant found to be associated with the trait and typed in the families of interest. The tested locus is assumed to be in strong linkage with the markers in the linkage peak. A disease-causing locus is defined as a functional site that affects the outcome of a disease but is not necessarily fully responsible for the disease. Assuming there are some disease-causing loci contributing to the linkage peak, our null hypothesis (H_0) is that the tested locus is in linkage equilibrium with any of the disease-causing loci. The alternative hypothesis (H_a) is that there is linkage disequilibrium between the tested locus and a disease-causing locus. Linkage between the tested locus and a disease-causing locus is allowed under either hypothesis.

We can think of each family in the study potentially representing a different population. Ascertainment, family structure, whether a disease-causing variant is segregating and if so which variant, and some other factors determine the allele frequencies for cases and for controls at the tested locus for a given family. For example, three families may be ascertained for having two or more cases among first or second degree relatives. The first family may consist of an affected sib-pair and may be segregating variant A at locus B. The second family may consist of three sibs of which two are affected and may be segregating variant C at locus D instead of variant A at locus B. The third family may

consist of a pair of grandparents and two grandchildren of which the grandmother and one grandchild are affected, and may not be segregating any disease variant at any disease locus, with the cases due to environmental factors.

Let i index over families. We define $\hat{f}_i$ as the estimate of variant frequency at the tested locus in cases from the population represented by family i, LOD_i as the estimate of the LOD score for this family (at the location of the linkage peak for the whole data set), and IBD_i as the true pattern of IBD sharing in the family at the tested locus respectively. Under the null hypothesis when there is no LD between the tested locus and a disease-causing locus contributing to the linkage peak, we require $\hat{f}_i$ to be chosen so that it is uncorrelated under the null hypothesis with LOD_i conditional on IBD_i. Also, the expected value of $\hat{f}_i$ must not depend on IBD_i under null hypothesis. We can achieve these requirements as follows.

First, we define two sets of markers from the available genotyped markers such that the markers in the first set (M1) are in linkage equilibrium with markers in the second set (M2). The first set, M1, are the ones used to calculate the LOD scores and determine the location of the linkage peak. The second set M2 should include the locus of interest that is to be tested for linkage disequilibrium with a disease-causing locus contributing to the linkage peak. They will be used to estimate the family variant frequencies. The tested locus should be in strong linkage but negligible linkage disequilibrium with some of the markers in M1. Selection of suitable markers for genotyping for two such marker sets is

feasible because linkage extends over a large scale compared to linkage disequilibrium. On the other hand, the assumption of independence between the two marker sets may be relaxed provided the tested locus doesn't contribute to the LOD score. For multipoint linkage analysis, family LOD scores over a chromosome region of interest will not vary significantly by the inclusion or exclusion of some markers if the remaining markers provide enough information about linkage over the given region.

A suitable variant frequency estimator for each family then needs to be defined. Any estimator of the form of weighted average of variant counts will do. Here we use the weighted average of variant counts over affected individuals within each family as the estimator of the variant frequency for a family. Since we are assuming no inbreeding and the tested locus does not contribute to the linkage peak, under H_0, it can be shown that the expected value of variant count conditional on IBD_i for a family is the expected variant frequency for the family. Thus, for the variant frequency estimator in weighted average form, under the null hypothesis, we can further prove that the second requirement that the expected value of $\hat{f}_i$ must not depend on IBD_i under null hypothesis is valid too.

Then the family LOD score and variant frequency estimate are uncorrelated if there is no linkage disequilibrium between the tested locus and a disease-causing locus contribution to the linkage peak (see Browning and Li, 2004 for detailed proof). At the disease-causing locus which is contributing to the LOD score peak, families with high LOD scores will tend to have more genetically caused cases, so that the family variant

frequency at the disease-causing locus will be correlated with the family LOD score. For a tested locus that is in linkage disequilibrium with the disease causing locus as for the alternative hypothesis, the estimated family variant frequency is correlated with the frequency at the disease-causing locus and thus also with the family LOD score. The null hypothesis and alternative hypothesis can be rephrased as "no correlation between family LOD score and variant frequency" and "there is correlation between family LOD score and variant frequency" at the tested locus.

Weighting Schemes

There are multiple valid possibilities for the weights. As long as the weights are independent of the observed genotype at the tested locus of the individual, the weighted average of variant counts is an unbiased estimate of the family frequency. Setting $w_{ij} = 1$ for all j yields a naïve estimator in which we treat cases equally and simply count alleles in them (WE, Weighted Equally). Alternatively, we may use weights based on prior IBD probabilities (given the family structure but not the genotypes) as in Browning (2004). We can emphasize alleles shared among affected individuals by weighting cases according to their kinship coefficient directly (WKS, Weighted by Kinship Shared). Let $kinship(j, j')$ be the kinship coefficient between individual j and j'. Here is the definition for Weighted by Kinship Shared scheme for case j in family i:

$$w_{ij} = \sum_{j'=1}^{n_i} kinship(j, j')$$

Since excess IBD sharing is the basis for nonparametric linkage analysis, we would expect WKS to perform better than the other weighting schemes. If there is marker and

disease association, families with high LOD scores will cosegregate the same allele/genotype. The same families will have higher family frequency estimates for the cosegregated allele/genotype using WKS than WE. Conversely, for those families with low IBD sharing or LOD scores, the family frequency estimates are close to each other by these two methods as affected individuals tend to have similar weights. However, the performance of the WKS method can be affected by other factors such as pedigree structure. If affected cases within a family are less correlated, i.e. low kinship coefficient, or all pairs of cases have the same kinship coefficient, such as sibs, or there are only two affected individuals per family, the family frequency estimate using WKS will be almost or exactly the same as with WE. Since the performance of WKS will be at least as good as WE, we ignore the WE method and only use WKS method in the remainder of this paper.

Test statistic

Spearman rank correlation coefficient (ρ) is used to measure the correlation between family LOD score and frequency. It is a nonparametric statistic based on ranking of family LOD score and frequency without making any assumptions about the distributions of the two variables. This avoids the model assumption violation and outlier problems involved in using Pearson correlation. Since the selection of the tested allele or genotype is random, we can obtain a positive or negative correlation if the disease associated one is selected or unselected. A two-sided hypothesis test can be constructed by applying the large-sample approximation (Hollander and Wolfe, 1999).

When there are only two categories (either two alleles or genotypes) for the tested locus, either category will give the same test result although they have opposite correlation coefficients. The choice of tested category doesn't matter under such situations. However, if more than two alleles or genotypes exist, the tests for correlation between family LOD score and frequency for different alleles or genotypes will give different results. Under such a situation, we test for correlation between family LOD score and each allele or genotype respectively and use the maximum test statistic or minimum p-value as an overall statistic for the tested locus. We denote this as "Extreme Statistic Method". Note that the correlation coefficients themselves are not comparable with each other. Their values depend on the data and carry no information about the correlation of the variables (except $\rho = 0$ indicates no correlation). Two uncorrelated variables can result in a correlation coefficient far from 0, while a correlation coefficient close to 0 may come from two highly correlated variables. On the other hand, test statistics or p values for correlation tests using different data are under the same distribution under null hypothesis such that we can use the maximum test statistic or minimum p value to make inferences. To evaluate the significance level of the overall statistic (maximum test statistic or minimum p-value), a permutation procedure is applied. To maintain the dependency among frequency estimates for all alleles/genotypes, family LOD scores are shuffled among the families under the null hypothesis and the frequency estimates are kept unchanged. For each permutation, a new overall statistic (maximum test statistic or minimum p-value) is obtained by performing the Spearman rank correlation tests again. After N permutations, we can get the approximate null distribution of the overall statistic

for the tested locus. The empirical p value for the tested locus is the proportion of the permuted overall statistics that are more extreme than the observed one.

We define three test types: allelic, genotypic and combined (allelic and genotypic) tests. The combined test uses the extreme statistic considering allele frequencies and genotype frequencies together.

Assuming a diallelic locus with alleles A and a, there are three possible genotypes, AA, Aa and aa. Testing correlation of family LOD score with the frequency of a genotype actually simultaneously tests the correlation between family LOD score and the frequency of the group of two remaining genotypes. If the disease is dominant and there is marker-disease association (high penetrances for AA and Aa), the correlation test statistic between family LOD score and frequency estimate of aa (i.e. the group of AA and Aa) will be the most extreme one and chosen as the overall statistic for the locus. If an allelic test is used for this model, the correlation signal will be weaker as both alleles contribute to the disease. It is the same for additive model that a genotypic test will outperform the allelic test. For a recessive disease model, genotypic or allelic tests should have similar power. Since the combined test considers allelic and genotypic together, its performance should always be close to the more powerful one of the allelic and genotypic tests.

Simulation study

A simulation study is carried out to evaluate the validity and power of the proposed method. We assume a disease with population prevalence of 0.1 and phenocopy rate of 0.3. The disease locus is assumed to have 2 alleles (*D/d*) with the first one (*D*) to be the disease causal allele. We vary the disease causal allele frequency from 0.1 to 0.9 with increment of 0.1. Let f_{DD}, f_{Dd} and f_{dd} be the penetrances of the 3 genotypes respectively. Under the constraint such that $f_{DD} = f_{Dd} > f_{dd}$ for a dominant model, $f_{Dd} = \frac{1}{2}(f_{DD} + f_{dd})$ and $f_{DD} > f_{Dd} > f_{dd}$ for an additive model or $f_{DD} > f_{Dd} = f_{dd}$ for a recessive model, four dominant models (disease causal allele frequency = 0.1, 0.2, 0.3, 0.4), four additive models (disease causal allele frequency = 0.1, 0.2, 0.3 or 0.4) and three recessive disease models (disease causal allele frequency = 0.3, 0.35 and 0.4) are constructed.

Besides the disease locus which is omitted from the analysis, two sets of additional markers are simulated. Markers in the first set are used for family LOD calculation and in linkage equilibrium with each other, the disease locus and markers in the other set. We simulate four markers with four equally frequent alleles for the first marker set. The distance between adjacent two markers is 2 cM. The disease locus is placed in the middle of the four markers. Markers in set two are tested markers in which we are interested. They are tightly linked to the disease locus. We simulate only one diallelic locus in nearly complete linkage with the disease locus (0.0001 cM). The frequency of the allele for the tested marker that is associated with the disease variant is assumed to be 0.1, 0.3, 0.5, 0.7 or 0.9. Linkage disequilibrium (D' or r^2) between disease and tested loci is set to be zero

for the null hypothesis and greater than zero for the alternative hypothesis. It can be shown that if the alleles at the tested and disease loci that are positively associated with each other have the same frequencies, r^2 is the square of D'.

We simulate 250 three-generation pedigrees with at least two affected individuals. First, a three-generation pedigree structure is randomly generated based on a multinomial distribution such that a parent has a probability of 0.2 of having one offspring, 0.5 for two offspring and 0.3 for three offspring. A traditional forward (from parent to offspring) simulation method is used to simulate genotypes and phenotypes from grandparents down to grandchildren for a three generation pedigree. For a complex disease with moderate genetic effect, a large proportion of simulated pedigrees will have no affected individuals and will be ignored, which wastes most of the simulation time. In order to improve the efficiency, we use a backward-forward simulation procedure. Here, "backward" refers to "from offspring to parent" as the segregation process from parents to offspring is reversible. One individual in a pedigree is chosen randomly as an index case. Conditional on his/her phenotype (affected), we generate his/her multilocus genotype. Then following the pedigree structure either up to the index case's parent or down to his/her offspring if applicable, genotypes of the parent or offspring are generated conditional on genotyped (genotype has been simulated) individuals at that time in the pedigree assuming Mendel's inheritance model and Haldane's map function. If the individual is a founder and no other correlated genotyped individuals at the time, his/her genotype is sampled from the population frequency distribution. Finally, the phenotype of the current individual is assigned according to his/her genotype. This process is done

recursively until genotypes and phenotypes of all individuals are assigned. Families with only one affected individual are discarded as they are not informative in linkage analysis. The disease locus is omitted and 5% of the genotypes over markers are randomly set to be missing before the analysis. Figure 4.1 list three examples of simulated pedigrees.

Results

Estimated type I error rate

To estimate the type I error rate under the null hypothesis, linkage disequilibrium between disease and tested locus is set to zero in the simulation. We test a variety of combinations of disease models and allele frequencies at 0.05 significant level. For each simulated data set, allelic, genotypic and combined tests are applied with the WKS weighting scheme. Table 4.1 gives part of results for estimated type I error. The standard errors for the estimated type I error rates over 1000 simulations are approximately 0.007. From the table, we see there is no estimated type I error rate significantly differ from 0.05 nominal level. The remaining data have similar results (not shown) except for one case when disease allele frequency equals 0.3, penetrances are 0.492, 0.061, 0.061 and the tested marker's allele frequency (for the allele in LD with the causal allele) is 0.1, the estimated type I error rates are 0.07~0.078 for the allelic, genotypic and combined tests. Further studies exploring the reasons for this observation are ongoing.

(Table 4.1 will be placed here)

Power study

To estimate power, a disease model needs to be specified which is difficult for complex diseases. Type 2 Diabetes, for example, has a population prevalence of ~5% (Raffle et al, 1996) and a phenocopy rate of 30~50%. We use 10% as the population prevalence and 30% as the phenocopy rate to construct disease models with various causal disease allele frequencies, assuming perfect LD between tested marker and disease locus. Figure 4.2

shows effect of causal disease allele frequency under additive, dominant and recessive disease models. For a fixed population prevalence and phenocopy rate, increasing disease causal allele frequency results in decreasing the relative effect of the disease causal allele/genotype(s). This makes the association with disease hard to detect. In figure 4.2, for additive, dominant and recessive disease models, the power of the association explaining linkage method we propose decreases as disease causal allele frequency increases.

(Figure 4.2 will be placed here)

In the previous simulations for the power study, linkage disequilibrium is set to be perfect between the tested and disease loci such that D' or r^2 is 1 and both loci have the same allele frequencies. In order to evaluate the effect of the allele frequency of the tested marker, we vary the allele frequency of the marker but keep perfect LD. The first marker allele is the one associated with the disease causal allele. Figure 4.3 shows the result of a dominant and recessive model. For the dominant model, the test is most powerful when the marker allele frequency is 0.1 which is equal to the disease causal allele frequency. For the recessive model, the disease causal allele frequency is 0.3 and the power curve of the test has a peak at the same marker allele frequency of 0.3. For other models not shown here, all support the same conclusion that, given complete LD between tested and disease loci, the test is most powerful when the tested and disease loci have the same allele frequency distribution.

(Figure 4.3 will be placed here)

D' and r^2 are measures of association between two loci. If we vary D' or r^2 between 1 and 0 but keep the same allele frequency distribution at the tested and disease loci, we can study the effect of LD. For D' values of 1, 0.8, 0.6, 0.4, 0.2 and 0, the corresponding r^2 are 1, 0.64, 0.36, 0.16, 0.04 and 0. As expected, the larger the value of D' *or* r^2, which means stronger association, the more powerful the test (shown in figure 4.4).

(Figure 4.4 will be placed here)

The choice of allelic or genotypic test depends on the underlying disease model. For additive and dominant models, the genotypic test is superior to the allelic one. This is mainly due to the fact that the "Extreme Statistic Method" we proposed for multi categorical data (>two genotypes) detects the maximum correlation between family LOD score and frequency of a group of two genotypes that contribute the most to the disease. These two genotypes include one homozygous and one heterozygous, which adds noise for the allelic test. However, when considering recessive disease, both allelic and genotypic test detect the signal with a single homozygous genotype or allele. Hence, allelic and genotypic tests for recessive disease models have similar power to each other. Results in figure 4.2, 4.3 and 4.4 confirm the above conclusion. For real data we don't know the underlying disease model, but the genotypic test should be a better starting point as it fits different models in general. Or we can use the combined test as it uses the extreme statistic from the allelic and genotypic tests together. We can see from figure 4.2, 4.3 and 4.4 that the results from the combined test are close to the more powerful of the allelic and genotypic tests.

Discussion

Non-parametric linkage analysis detects the chromosomal regions in which there is excess IBD sharing among cases in a pedigree. However, different pedigrees that contribute to a linkage signal might share different alleles IBD in excess as the excess IBD sharing does not refer to a specific variant. Association methods test for the difference of frequencies of a specific variant between cases and controls. By asking whether an association can help to explain a linkage result, these two approaches can be combined together. We propose a new method for testing association explaining non-parametric linkage for general pedigrees. The correlation between family LOD score and allele/genotype/haplotype frequency is tested to address whether the association partially explains the linkage result. It provides a way to distinguish false from true positive association signals. On the other hand, it also enables us to accept weak association/linkage signals given the fact that our test is positive.

Our method has been proved to have a valid type I error rate in most of tested disease models by simulations studies. To gain optimal power, there are several key issues. The first is to select the most informative markers for linkage analysis. This is not very difficult to achieve since most linkage studies are done by multipoint analysis with markers with multiple alleles. The second issue is selection of a weighting scheme. Weighted by Kinship Shared (WKS) is the more powerful weighting method as it emphasis alleles/genotypes shared among cases for extended pedigrees. However, if most pedigrees have only a pair of cases or IBD sharing among affected relatives are almost the same (like sibs), Weighted Equally (WE) can be a good alternative as it gives similar

frequency estimates to the WKS weighing scheme. The third issue is the choice of test type. For recessive disease model, there should not be any significant difference between allelic and genotypic tests. However, for additive and dominant disease model, different tests result in a difference of power. Genotypic tests are more powerful than allelic tests, since the test we proposed for multiple categorical situations will detect the strongest association between family LOD score and frequency estimate of the group of the two genotypes contributing the most to the disease. The combined test includes information from both allelic and genotypic tests such that its performance is as good as the more powerful one. In general, the combined test or genotypic test should be chosen if the underlying disease model is unknown. Finally, and most importantly, the last key issue is how to select tested markers. The association between a marker and disease locus depends on both allele frequencies and distance between them. The location of the disease variant is of the primary interest. The allele frequencies of the tested and disease loci have a big impact on the power of the test we proposed. The power is the highest when they have a matched allele frequency. However, as we mentioned before, the underlying disease causal allele frequency is normally unknown. This makes the choice of markers very difficult.

In real data analysis, it might be convenient to test markers used to estimate the LOD score in the proposed test. This violates the assumption of independence of the two sets of markers used for linkage and association explaining linkage tests. However, family LOD scores over a chromosome region of interest will not be affected significantly by some markers as long as the remaining marker set provides enough information about

linkage over the given region. Therefore, these dependent markers will provide a feasible test. But we suggest including some independent markers, especially in the flanking area to make family LOD estimation more accurate.

In a recent paper by Li et al. (2004), the authors proposed a new test GIST assessing whether an allele can account in part for a linkage signal. They assigned weights to families on the basis of the genotypes of all cases in the same families and test for correlation between family LOD score and weight. The weight variable in their method is comparable to the estimated family frequency in our method. As the weights didn't take relationships among cases into account, they are similar to one of the weighting method we propose, Weighted Equally (WE). The additive weighting scheme in their paper is the same as our allelic test by WE. For the other two weighting schemes (dominant and recessive) they used, there are no corresponding weighting and testing strategies by our method. For the additive weighting scheme they used, the authors concluded it should perform the best under an additive disease model. As we can see in our results, the allelic test is not necessarily the most powerful one for an additive disease model.

Although we construct the test based on non-parametric linkage results, the test can be applied to parametric linkage results. For non-parametric linkage analysis, we can obtain family LOD score under the disease model that shows the strongest linkage and test for its correlation to the family frequency estimate.

It has been shown that the affected individuals are not only likely to share alleles at a single locus, but also at the surrounding haplotype (Van der Meulen and Te Meerman, 1997). Although we present results only for single point allelic, genotypic and combined tests, it is possible to extend to a haplotype based test. For phase known data set, we can simply replace the allele/ genotype count by haplotype count and conduct the test using the multi-category procedure. If phase is unknown, since the tested markers are usually tightly linked, it is reasonable to assume no recombination during haplotype assignment. Each pedigree might have several possible haplotype configurations. Family haplotype frequency estimate should be a weighted average over all possible configurations. Haplotype frequencies need to be estimated first to assign weights for those configurations. The combined test then can be extended to include both single point and haplotypic tests together by using the extreme statistic from all of the test statistics available.

References

Browning, S.R. (2004) Case-control single-marker and haplotypic association analysis of pedigree data. *Genet Epidemiol. (in press)*

Browning, S.R. and Li Li (2004) Association explaining non-parametric linkage for binary traits in general pedigrees. *in preparation.*

Hollander M and Wolfe D. A (1999) *Nonparametric Statistical Methods*, 2nd edition, New York, Wiley

Horikawa, Y. et al. (2000) Genetic Variation in the Gene Encoding Calpain-10 is Associated with Type 2 Diabetes Mellitus. *Nature Genetics 26: 163-175*

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

Li, C. Scott, L. J. and Boehnke, M. (2004) Assessing Whether an Allele Can Account in Part for a Linkage Signal: the Genotype-IBD sharing Test (GIST) *Am. J. Hum. Genet. 74:418-431*

Raffel, L., Robbins, D. et al. The GENNID study: a resource for mapping the genes that cause NIDDIM. *Diabetes Care* 19:864-872

Sun, L., Cox, N. J. and McPeek, M. S. (2002) A Statistical Method for Identification of Polymorphisms That Explain a Linkage Result. *Am. J. Hum. Genet. 70:399-411*

Van der Meulen MA, te Meerman GJ. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genet Epidemiol* 14:915-920, 1997.

Table 4.1. Estimated type I error at 0.05 significant level over 1000 simulations.

Minor allele frequency of tested locus equals disease causal allele frequency and allele frequencies of marker for LOD score are 0.5 and 0.5. The standard errors for the estimated type I error rates over 1000 simulations are approximately 0.007.

Disease Model	Penetrance	Allelic	Genotypic	Combined
Dominant				
DAF = 0.1	(0.368,0.368,0.037)	0.064	0.054	0.053
DAF = 0.3	(0.137,0.137,0.061)	0.062	0.063	0.055
Additive				
DAF = 0.1	(0.666,0.351,0.037)	0.058	0.056	0.054
DAF = 0.3	(0.190,0.125,0.061)	0.059	0.06	0.062
Recessive				
DAF = 0.3	(0.492,0.061,0.061)	0.058	0.052	0.06

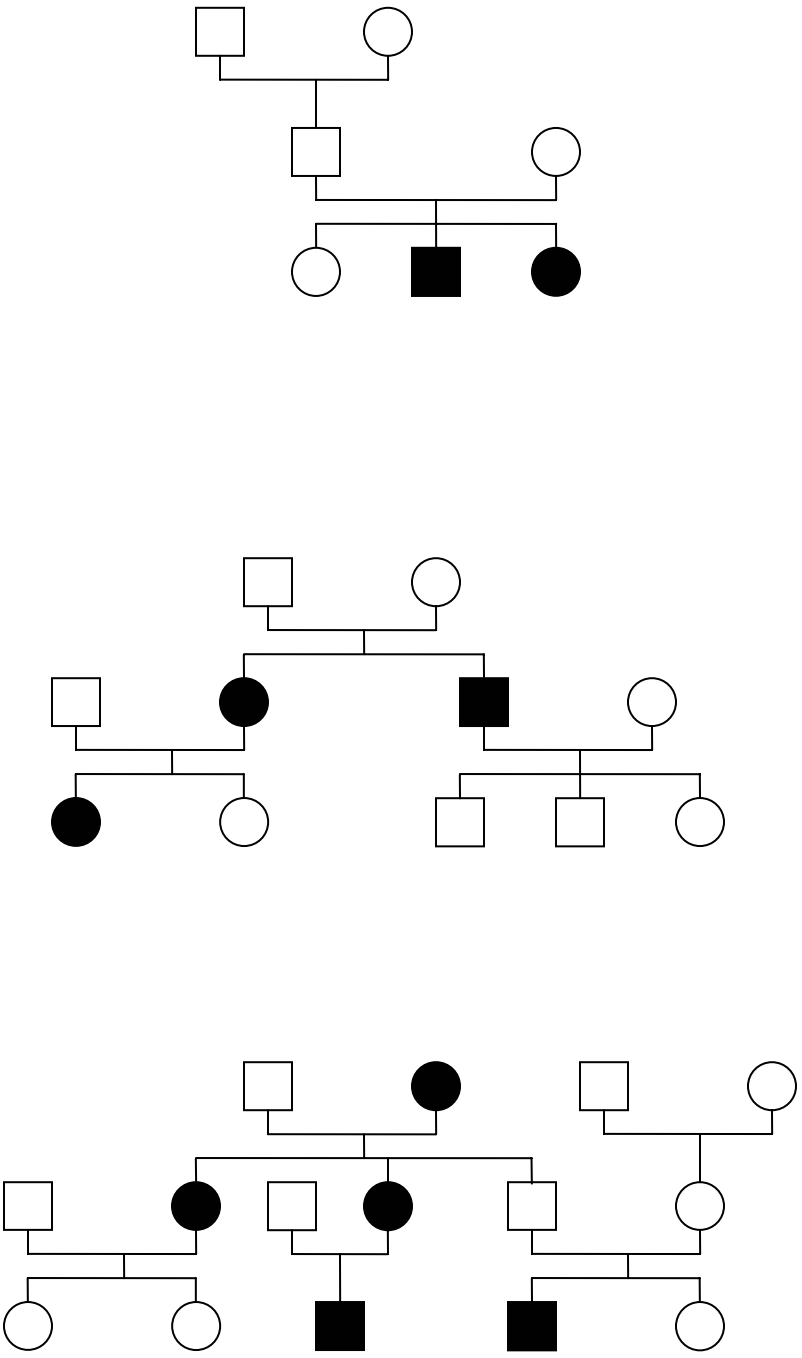
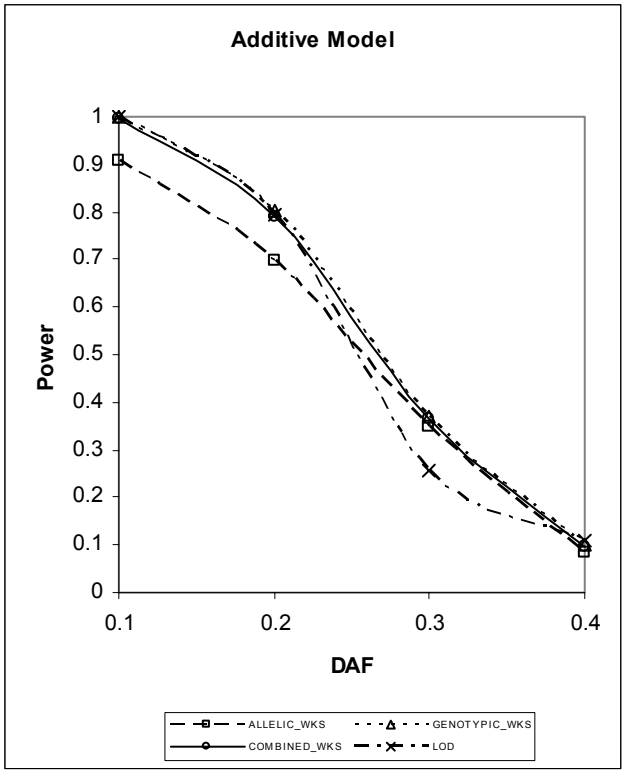


Figure 4.1. Examples of simulated 3 generation pedigrees.

Squares represents males, while circles represent females. Shading indicates affected cases.



(a) additive disease model

Figure 4.2 Effect of disease allele frequency.

Disease population prevalence is 0.1 and phenocopy rate is 0.3. Marker for LOD score has two equally frequent alleles. The tested and disease loci are in perfect LD (same allele frequency distribution and complete LD).

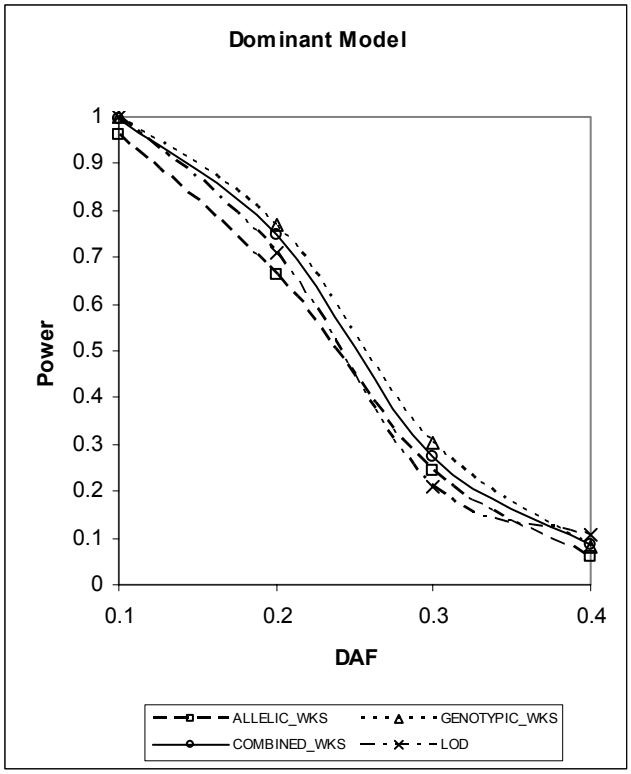


Figure 4.2 (b) dominant disease model

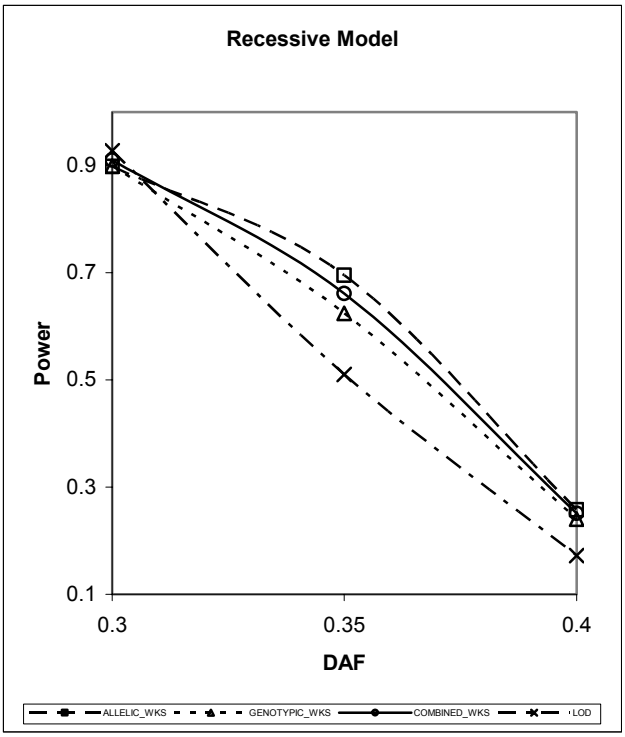
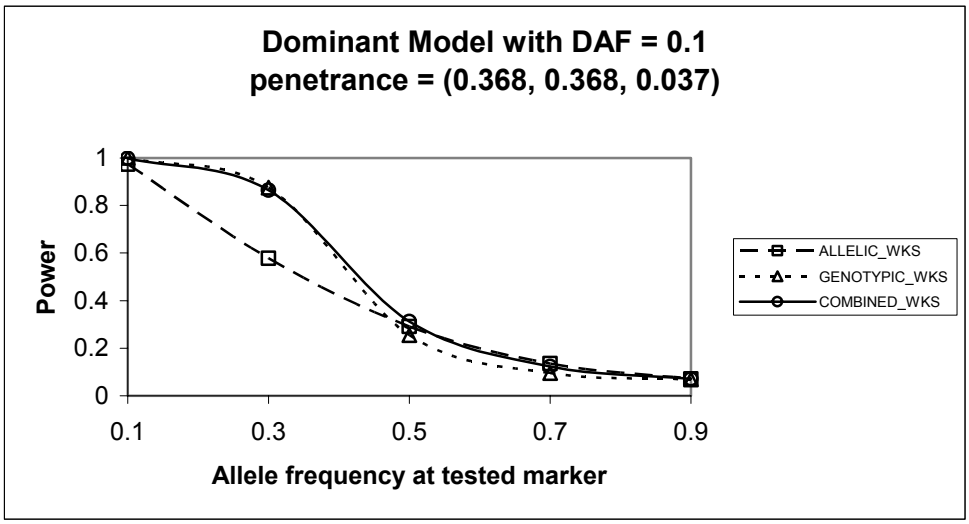
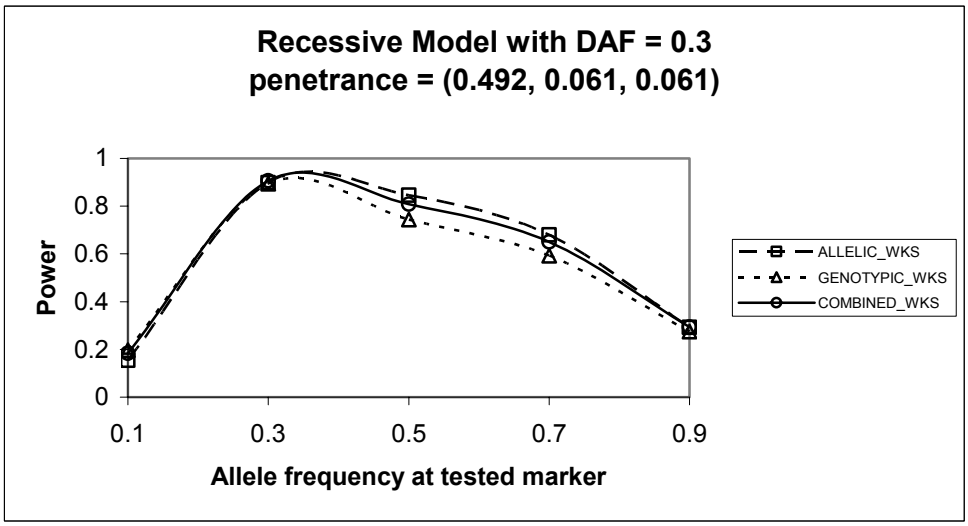


Figure 4.2 (c) recessive disease model.



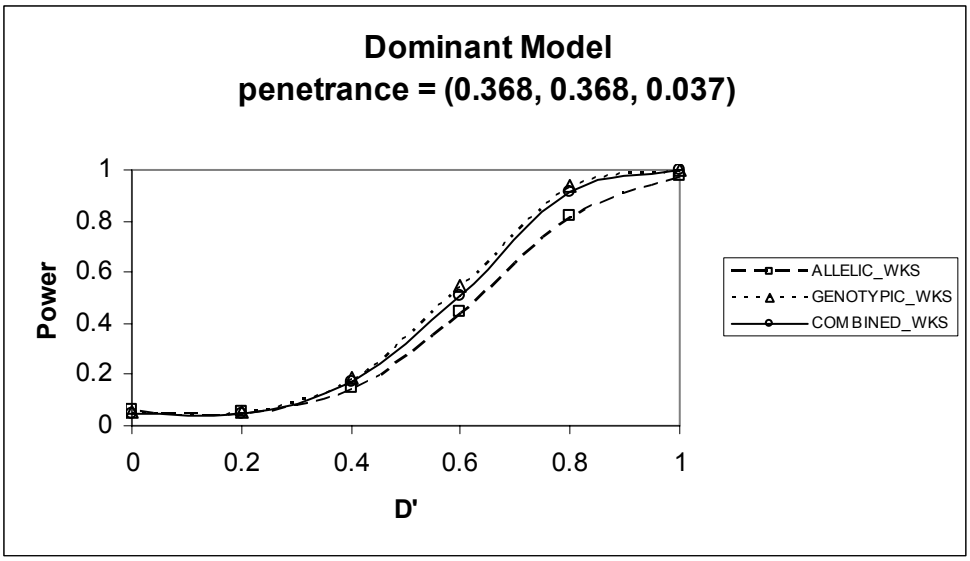
(a) dominant disease model



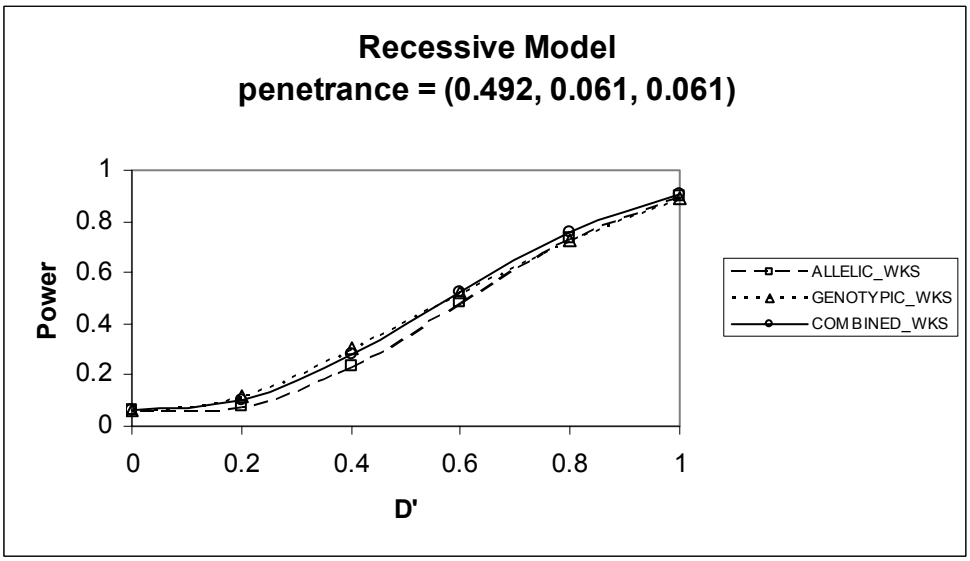
(b) recessive disease model

Figure 4.3. Effect of allele frequency of tested locus.

Disease population prevalence is 0.1 and phenocopy rate is 0.3. The tested and disease loci are in complete LD.



(a) dominant disease model



(b) recessive disease model

Figure 4.4. Effect of *D′*.

Disease population prevalence is 0.1 and phenocopy rate is 0.3. The tested and disease loci have the same allele frequency.