INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

STUDY DESIGN ISSUES IN THE ANALYSIS OF COMPLEX GENETIC TRAITS

by

Katrina Blouke Goddard

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

1999

Approved by Ellen M. Wijsman
Chairperson of Supervisory Committee

Program Authorized to Offer Degree Biostatistics

Date Feb. 16, 1999

Doctoral Dissertation

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, P.O. Box 975, Ann Arbor, MI 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Katrina Goddard

Date 2/16/99

University of Washington

Abstract

Study Design Issues in the Analysis of Complex Genetic Traits

by Katrina Blouke Goddard

Chairperson of the Supervisory Committee:
Professor Ellen M. Wijsman
Department of Biostatistics and Division of Medical Genetics, Department of Medicine

Many common diseases that potentially have a large public health impact, such as heart disease, cancer, and diabetes, are known or thought to have a genetic component to risk. These traits are often complex with multiple contributing genetic and non-genetic factors. The identification of genetic risk factors through genetic mapping and cloning studies may aid our understanding of the underlying mechanism of disease. Unfortunately, large sample sizes are often required for linkage analysis of complex traits, resulting in a need to identify cost-effective strategies. The use of power and sample size considerations as the criteria for evaluating methods may grossly underestimate the utility of some designs. Additional factors can be incorporated into the analysis by considering the overall cost of a study including the pedigree collection, marker genotyping, and statistical analysis. Using analytical and simulation methods, we evaluate three study design issues relevant to cost-effective linkage analysis of complex traits.

First, we demonstrate that pedigrees of the same size and structure that differ in the number of affected and unaffected individuals also differ in the probability of segregating a particular trait locus, which affects the power to detect linkage at that locus. Although the optimal pedigrees for detecting linkage to a particular trait locus are highly dependent on the trait model, we identify pedigree selection schemes that are cost-effective for a vari-

ety of underlying trait models. Second, we identify characteristics of single nucleotide polymorphisms (SNPs) for a cost-effective genome screen compared to the current microsatellite-based technology. Issues that are addressed in comparing the markers include the information content for clustered SNPs in the presence or absence of linkage disequilibrium, the marker spacing, and the map accuracy. Finally, we describe a method called 'downcoding' to reduce the computational burden associated with currently available likelihood-based methods that are potentially more powerful for detecting linkage than alternative methods. It is often not possible to use these methods without downcoding because of limitations in the available computational resources. Through consideration of the three topics described above, we are able to identify cost-effective strategies for the analysis of complex genetic traits.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I owe special thanks to my advisor, Ellen Wijsman, for her guidance and support while I have been a graduate student. I would also like to thank my committee members for their helpful comments and interest, Barbara McKnight, Elizabeth Thompson, Phil Green, and Scott Davis. Finally, I wish to thank my fellow students for their encouragement, especially James Lovato, Jinko Graham, Aparna Anderson, and Carl DeMoor. This work was supported in part by NIH grants AG05136 and 5T32CA09168.

DEDICATION

The author wishes to dedicate this dissertation to her husband Shawn Goddard, and to her parents Morley and Kay Blouke. Thank you for your patience and support.

CHAPTER 1 : INTRODUCTION

Genetic studies are an important tool for elucidating the underlying biological mechanism of disease. Identification of the genetic risk factors that influence the development of disease is a first step towards understanding the role of these factors in the causation or prevention of disease. In addition, understanding the role of genetic risk factors provides a framework for understanding non-genetic forms of disease, as well as understanding the normal function of genes. Benefits of identifying genetic risk factors include determining the carrier status of individuals through genetic testing, understanding the role of genes in drug performance for the treatment of disease, and preventing or treating disease through the development of pharmaceutical agents.

The process of identifying genes is divided into two main steps: genetic mapping, and positional cloning. Genetic mapping is the process of identifying a small region of a chromosome that contains the gene of interest. Positional cloning is the process of physically isolating the relevant gene, which allows us to identify the exact genetic defect that increases the risk of developing the disease. Here we focus on issues related to the process of genetic mapping, and we do not consider the process of positional cloning.

Several tools have been developed for conducting genetic studies in a systematic manner. Genetic markers are one of the molecular tools that aid us in this process. A genetic marker is a locus with a known location in the genome. For genetic mapping studies, we usually require that the locus has variation at the population level (i.e. there are at least two variants where the most common variant has frequency < 0.99). Genetic mapping proceeds by identifying a marker that is physically located near the gene of interest, or 'linked' to the gene. A genetic map is a set of markers that all have a known location. Genetic maps are characterized by the number of markers on the map and the estimated distances between the markers. Statistical methods are used to determine if there is statistically significant evidence for linkage between a marker and the disease locus. The lod

score method (Morton, 1955) is one example of a statistical test that has been developed for this process. Initially, the computations were either performed by hand, or obtained from published tables for a few simple pedigree structures (Morton, 1955). However, computer algorithms have been developed to allow larger, more complex problems to be considered than what can be easily computed by hand. For example, Ott (1974) implemented the lod score method in a computer program that could evaluate general pedigree structures.

Genetic mapping and cloning has been an enormously successful method for identifying genes that contribute to single-gene disorders. The first disease locus to be mapped using DNA markers was Huntington's disease (Gusella et al., 1983). Since then, approximately 1792 disease genes have been mapped to a chromosomal region (http://www3.ncbi.nlm.nih.gov/Omim), and 108 disease genes have been identified using positional cloning methods (http://genome.nhgri.nih.gov/clone). Unfortunately, studies involving complex traits are often not as successful. One important characteristic of complex traits that may contribute to the poor performance of these methods is that there are multiple factors that influence the risk of developing disease. Numerous studies have demonstrated that the power to detect linkage is reduced when heterogeneity is present in the sample (e.g. Risch, 1990; Goldin and Weeks, 1993; Goldin and Gershon, 1988). It may be difficult to obtain the larger sample sizes that are required because of the decrease in power. In addition, for analysis methods that require the specification of model parameters, the power to detect linkage is reduced if the model parameters are misspecified (Clerget-Darpoux et al., 1986; Vieland et al., 1992a,b; Greenberg and Hodge, 1989). For complex traits with multiple factors influencing the trait, it is often very difficult to accurately estimate the model parameters.

Many diseases that have a large public health impact are complex disorders. Some examples of common diseases that are known or thought to have a genetic component to the risk include cancer, Alzheimer's disease, diabetes, and heart disease. In the United States, heart disease and cancer were the first and second leading causes of death respectively in

1992 (Parker et al., 1996). Diabetes was the seventh leading cause of death in this time period. The lifetime risk of developing breast cancer is 1 in 8 for women, while the lifetime risk of developing prostate cancer is 1 in 5 for men. The lifetime risk of developing colorectal cancer is approximately 1 in 16, with a slightly lower risk for females than for males.

Because of the large public health impact of complex diseases, it is important to identify strategies that are efficient for the study of complex traits. There are several aspects of study design that can be evaluated to identify effective strategies. First, we have a choice in the selection of individuals for inclusion in the analysis. The possibilities range from independent individuals, to large multi-generation pedigrees. Second, there are many different types of markers that can be used for genetic studies. The efficiency of the study design may depend on the choice of marker type and the structure of the genetic map including the number of markers, and the marker spacing that is used. Finally, we must also consider the different types of statistical analysis that are available. In this chapter we review many of the types of markers and the statistical models and methods that are available for genetic studies. We also describe three study design issues that are studied in this dissertation. Each chapter addresses a different issue in study design including the selection of pedigrees, the choice of markers, and the reduction of computation time for likelihood-based linkage analysis.

1.1 MOLECULAR TOOLS AVAILABLE FOR GENETIC MAPPING

Genetic markers are the molecular tools that are necessary for genetic mapping. Here we describe different types of genetic markers that are available. The markers are compared in terms of three important characteristics for genetic mapping that are defined below.

One important characteristic of markers is the marker polymorphism. A locus is polymorphic if there exists more than one variant (or *allele*) at the locus in the population. The number of alleles present for a marker ranges from two up to approximately 15-20

alleles or more. The polymorphic information content (PIC) is a measure of the marker polymorphism that takes into account the number of alleles, and the allele frequencies (Botstein *et al.*, 1980). If a marker is linked to a disease gene, the marker alleles are transmitted from parent to offspring along with the disease phenotype because of the close proximity of the marker and the disease locus on the chromosome. In order to determine which marker allele is transmitted from a parent, the marker alleles for that parent must be different, and both parents and the offspring can not all have the same genotype. We are able to determine which parental allele is transmitted to the offspring more frequently when the marker has a high PIC. Therefore, markers with a high PIC are more useful for detecting linkage between a marker and a disease gene.

A second important feature is the presence of the marker type throughout the genome. In order to ensure that there is always a marker near the region of interest, we must be able to identify markers throughout the genome. Ideally, there should not be any regions of the genome where it is not possible to identify the markers, and it should be possible to identify multiple markers in a small region of the genome. The first genetic map of the human genome contained 403 markers (Donis-Keller et al., 1987). Genetic maps with more than 8000 highly informative markers are currently available, which is only a small proportion of all of the markers that have been identified (Broman et al., 1998).

Finally, the methods used to identify and detect markers should be easily extended to additional markers, and it is desirable to be able to automate the detection method. Most genetic studies require a large amount of genotyping. Automation can reduce the cost of genotyping, and increase the accuracy. Creating new methodology for each marker is very costly and labor intensive, so methods that can be used to identify and detect markers throughout the genome are more cost-effective.

The first markers were blood group and functional markers (Ott, 1991). These are protein molecules on the surface of the blood cells, where differences among individuals are detected using a variety of laboratory assays. These markers do not have many desirable

characteristics in that they generally have a low PIC, and a different assay must be developed for each marker. In addition, these markers are not generally found throughout the genome.

Blood group markers were replaced by markers called restriction fragment length polymorphisms (RFLP) (Botstein et al., 1980). Restriction enzymes are molecules that cut DNA at a particular sequence called a recognition site. A mutation or other change (*e.g.* insertion or deletion) in the recognition sequence eliminates the ability of the restriction enzyme to cleave at the recognition site. The presence or absence of the recognition site creates the variability that is used for RFLPs. Alternatively, some RFLPs are due to an insertion or deletion between two recognition sites. This type of polymorphism exists throughout the genome, so it is possible to screen the entire genome using these markers. However, RFLPs generally have a low PIC, since they only detect the presence or absence of a recognition site. In addition, the detection methods used for RFLPs are tedious, and require a large amount of sample DNA from each individual.

Variable number of tandem repeats (VNTR) (Nakamura et al., 1987) markers were used because they have a higher PIC than previous markers. The genome contains regions where a specific sequence of DNA is repeated a number of times. Variation is due to a difference in the number of repeats among individuals. The detection methods used for VNTR markers are based on the same laboratory methods that are used for RFLP genotyping, and thus are very labor intensive, and not easily automated. In addition, VNTRs are not uniformly spaced throughout the genome, but are clustered towards the telomeres.

Microsatellite markers, or short tandem repeat polymorphisms (STRP), (Litt and Luty, 1989; Weber and May, 1989) replaced VNTRs. These markers are conceptually the same as the VNTRs, but they differ in the size of the repeat unit. The repeat unit for STRPs is only 2-4 base pairs compared to 11-60 base pairs for the VNTR repeat unit. An important difference is that the detection method can be at least partially automated which allows the genotyping to be performed in a shorter period of time, and may improve the accu-

racy. It is possible to simultaneously genotype up to 15 STRPs compared to only one genotype in previous methods (Ziegle et al., 1992). STRPs generally have a high PIC, and are found throughout the genome.

A new type of marker has recently been developed called single nucleotide polymorphisms (SNP). These markers are based on differences in the DNA sequence at only a single base pair, and generally only have two alleles. Single base pair substitutions are found approximately every 1-2 kb, so a very large number of these markers are available throughout the genome. Detection methods are currently under development that could potentially be used to genotype hundreds or thousands of marker loci simultaneously in a completely automated system (Saiki et al., 1989; Wu et al., 1989; Nickerson et al., 1990; Pease *et al.*, 1994; Livak et al., 1995; Chee et al., 1996). An important disadvantage of using SNPs is that they have a low PIC since they are only diallelic systems. One possibility to increase the PIC is to combine multiple SNPs to create 'super-alleles' (Nickerson et al., 1992). If the markers are close enough together there is virtually no recombination between them, so the super-alleles are transmitted through the pedigree as if they are one locus. Another possibility is to use uniformly spaced SNPs combined with multipoint linkage analysis to extract information from neighboring markers.

1.2 STATISTICAL MODELS: SIMPLE VS. COMPLEX TRAITS

A simple genetic disorder is one where genetic defects at only one gene cause the disease with a simple mode of inheritance such as dominant, recessive, or codominant. These traits are also called mendelian disorders because they follow the laws of mendelian inheritance. Many genes that cause diseases that follow mendelian inheritance have been identified. Examples include Huntington's disease (Gusella et al., 1983; The Huntington's Disease Collaborative Research Group, 1993), and cystic fibrosis (Tsui et al., 1985; Rommens et al., 1989; Riordan et al., 1989; Kerem et al., 1989). Unfortunately, the majority of traits are not simple disorders, but have some complexity to the inheri-

tance such as multiple genetic and non-genetic forms of the disease. In this section, we outline several factors that contribute to a complex inheritance pattern.

The first type of complexity is *locus heterogeneity*. In this situation, there exists more than one locus that can cause the disease, such that each locus acts independently of the others. Inter-familial (between families) and intra-familial (within a family) heterogeneity reduces the power to detect linkage, since not all individuals are segregating the same disease locus. There are a large number of diseases that have been shown to have locus heterogeneity. One example is Alzheimer's disease. Three loci have been identified that cause the early-onset form of the disease on chromosomes 21 (Goate et al, 1991), 14 (Schellenberg et al, 1992), and 1 (Levy-Lahad et al, 1995). Breast cancer is a second example where two loci have been identified that contribute to early-onset familial breast cancer, BRCA1 (Hall et al, 1990; Miki et al., 1994) and BRCA2 (Wooster et al, 1994; Wooster et al., 1995).

A second form of complexity is *allele heterogeneity*. In this situation, different alleles at the locus cause the disease and may result in a different severity of disease, or different diseases altogether. Cystic fibrosis is one example where there are over 800 different mutations identified in the CF gene (http://www.genet.sickkids.on.ca/cftr-cgi-bin/MutationTable). Mutations that occur in different regions of this gene cause a range of phenotypes from severe forms of cystic fibrosis to male infertility in otherwise normal males (Kerem et al., 1996). A second example is the FGFR3 gene (Bonaventure et al., 1996). Mutations in different regions of this gene cause phenotypes with a different severity (some lethal) of short stature and reduced bone growth. In this case, these diseases were originally thought to be different diseases, rather than different forms of the same disease.

Penetrance is the probability of observing a particular phenotype given a genotype. This probability can range between the values of zero and one. *Reduced penetrance* is when the penetrance is some value in between the two extremes. In this case, there is more am-

biguity in the relationship between the phenotypes and the genotypes of individuals in the pedigree. Pedigrees are generally less informative in this situation since there is more unknown information. The penetrance may also depend on some other covariate such as age, diet, or another genetic locus that can add to the complexity of the model.

Finally, *multilocus inheritance* is a form of complexity where two or more loci act together to contribute to the disease risk. There are two general models that are commonly used to describe this type of inheritance, although there are other possible models that could be considered. *Polygenic* inheritance refers to a large number of loci that all contribute a small effect to cause the phenotype. We do not expect to be able to identify these loci because each effect is so small. *Oligogenic* inheritance refers to a situation with only a few loci that each contributes a large effect to the phenotype. There are two subgroups of these models that are defined by how the loci interact. In *additive* models the phenotype is determined by adding the effects of the individual loci. In *epistatic* models there are multiplicative interactions between the individual loci that determine the phenotype. In practice, we may not be able to distinguish the difference between a polygenic disorder and an oligogenic disorder. In fact, some disorders that we think of as having 'polygenic' inheritance may become 'oligogenic', as we are able to detect loci with smaller effects.

1.3 STATISTICAL METHODS USED IN THE ANALYSIS OF COMPLEX TRAITS

There are several different choices for the statistical analysis method. The methods differ in the type of samples required ranging from case-control studies where individuals are collected, to pedigree-based methods where families are collected. In addition, the amount of information that is required about the genetic model is different among methods. Some of the alternatives are outlined in this section, with the relative advantages and disadvantages of each method.

1.3.1 THE LOD SCORE METHOD

The lod score method has been successfully used to identify disease genes for many simple and complex disorders. Given observed data, the likelihood under the hypothesis of no linkage (H_0: $\theta = ½$) is compared to the likelihood under the hypothesis of linkage (H_A: $\theta < ½$). The lod score for a single family is then calculated as the base-10 log of the likelihood ratio (Morton, 1955):

$$LOD = \log_{10}\left\{\frac{L(\theta)}{L(1/2)}\right\},$$

where $L(\theta)$ is the likelihood under the model of linkage, and $L(1/2)$ is the likelihood under the model of no linkage. In order to calculate the likelihood, we must specify the model parameters for the inheritance of the trait. For the marker locus, we generally need to specify the number of alleles and the allele frequencies. For the trait, we generally assume there is a single trait locus with two alleles, D and d, and allele frequencies, p and $q = 1\text{-}p$, respectively. For a discrete trait the penetrance of each genotype is specified, which we parameterize as $f = \{f_{DD}, f_{Dd}, f_{dd}\}$. For example, for a fully penetrant dominant disorder with no sporadic cases, the penetrance is $f = \{1,1,0\}$, while for a fully penetrant recessive disorder the penetrance is $f = \{1,0,0\}$. For a dominant disorder with reduced penetrance, the penetrance is $f = \{a,a,0\}$, where $0 < a < 1$. For a quantitative trait the mean, μ_i, and the standard deviation, σ_i, are specified for each genotype (i = DD, Dd, or dd).

Early-onset breast cancer is one example where the lod score method has been successful in identifying two loci that contribute to this disease. BRCA1 is located on chromosome 17 (Hall et al., 1990; Miki et al., 1994), and may be responsible for familial breast cancer that is characterized by early-onset cases and a predisposition towards ovarian cancer. BRCA2 is located on chromosome 13 (Wooster et al., 1994; Wooster et al., 1995), and is found in families with early-onset disease and both male and female cases.

Colon cancer is a second example where the lod score method resulted in identification of multiple genes for two forms of disease. A gene that causes familial adenomatous polyposis (FAP) was localized to chromosome 5 (Bodmer et al., 1987; Leppert et al., 1987). In this form of colon cancer, patients develop hundreds or thousands of benign tumors in the colon. Cancer is caused by the progression of one of these tumors to a malignant form. In addition, two genes that increase the risk of hereditary nonpolyposis colorectal cancer (HNPCC) have been localized to regions on chromosome 2 (Peltomäki et al., 1993), and chromosome 3 (Lindblom et al., 1993). This latter form of the disease is characterized by familial clustering of colon cancer in addition to an early-onset of disease (<50 years) for at least one member of the family.

The examples presented above demonstrate that the lod score method can be successfully used to identify disease genes for complex traits. However, it is unclear whether this method will be successful in evaluating all types of complex disease. Numerous studies have demonstrated that the lod score method has reduced power when there is inter- or intra-familial heterogeneity. In each of the examples above, it was possible to identify a more homogeneous subset of the population that was segregating only a single form of the disease. The reduction in the heterogeneity of the sample contributed to the success of this method in these cases. A second limitation to the lod score method is that the power may also be reduced if the model parameters are misspecified (Clerget-Darpoux et al., 1986). Numerous studies have shown that the one-locus lod score method may be adequate for detecting linkage when the true underlying model is a two-locus model; however, it is very important to specify the mode of inheritance correctly (Greenberg and Hodge, 1989; Vieland et al., 1992a; Vieland et al., 1992b; Vieland et al., 1993; Goldin and Weeks, 1993). For complex traits, it is often difficult to estimate the model parameters accurately. Finally, the computation time necessary to perform parametric linkage analysis for many complex traits can be very high, especially for diseases where there is missing data, where a complex model must be specified, or if large sample sizes are required.

1.3.2 EXTENSIONS OF THE LOD SCORE METHOD

The lod score method has recently been extended to include two-trait-locus models. The likelihood of the data is calculated under a model with two genetic loci that cause the disease, locus A and locus B. Under this model, there are two alleles at each locus, with allele frequencies p_A, p_a, p_B, and p_b. There are nine penetrances that must be specified for the joint genotypes $f = \{f_{AABB}, f_{AABb}, f_{AAbb}, f_{AaBB}, f_{AaBb}, f_{Aabb}, f_{aaBB}, f_{aaBb}, f_{aabb}\}$. The power to detect linkage using a two-trait locus two-marker locus model may be significantly increased over the power using a one-locus model since the model can be more accurately specified (Shork et al., 1993). A disadvantage to using the two-trait-locus two-marker-locus model is that it can be computationally intensive, particularly if there are missing data. In addition, there are more tests that must be performed for this model since pairwise testing of n marker loci requires $n(n\text{-}1)$ analyses compared to only n analyses for methods that test one marker at a time.

A second extension to the lod score method is multipoint linkage analysis. In this method, multiple marker loci are considered simultaneously to increase the information available about recombination with the disease locus, which increases the power of the test. There are several disadvantages to using this approach. Multipoint linkage analysis is not as robust to model misspecification as single-marker linkage analysis (Risch and Giuffra, 1992). In single-marker linkage analysis, the effect of model misspecification is to inflate the recombination fraction, while the lod score is not greatly changed. However, in multipoint linkage analysis, it is not possible to inflate the recombination fraction when the disease locus is located between two flanking markers, so the lod score is reduced. A second disadvantage is that multipoint linkage analysis is computationally intensive. In current program implementations, there are restrictions either in the number of markers that can be considered simultaneously (e.g. VITESSE, O'Connell and Weeks, 1995), or in the pedigree size and structure (e.g. GENEHUNTER, Kruglyak et al., 1996).

1.3.3 SIB-PAIR APPROACH

One alternative to the lod score method is the sib-pair approach. There are several different sib-pair methods considered here, with similar advantages over the lod score method. It is attractive to consider methods using sib-pair data since sib-pairs are often easier and less expensive per individual to collect than more extended pedigrees. Although sib-pair data can also be analyzed using the lod score method, the sib-pair methods described here differ from the lod score method in that it is not necessary to specify many of the model parameters. As mentioned previously, it may be difficult to accurately specify the model parameters for many complex traits, so the sib-pair methods may be more powerful than lod score methods if the model parameters are misspecified. In addition, the computation time necessary to perform the analysis is often much less for sib-pair methods than for lod score methods, particularly for complex traits.

Penrose (1935) proposed the original sib-pair method. In this method, sibs are categorized as either being alike or not alike in phenotype at each of two loci (*i.e.* a marker and a trait). If linkage is present, sibs that are concordant for the trait phenotype tend to be alike at the marker locus and sibs that are discordant for the trait phenotype tend to be dissimilar at the marker locus. A standard chi-square test can be used to determine the significance of the deviation.

Often, only the affected sib-pairs (ASPs) are used in the analysis, since Suarez et al. (1978) showed that sib-pairs where both sibs are affected are often more informative for linkage. Under the hypothesis of no linkage, the probability that ASPs share i alleles identical by descent (IBD) at the marker locus is ¼, ½, ¼ for i = 0, 1, or 2 respectively. However, if linkage is present, there will be an excess of allele sharing over that expected under the hypothesis of no linkage. Several test statistics have been proposed to measure the significance of the excess sharing of alleles by ASPs. Blackwelder and Elston (1985) compared three test statistics, and found that the mean number of marker alleles shared

IBD had the best performance in terms of power and significance level for most of the models tested.

The Haseman and Elston (1972) approach is another sib-pair method that uses a regression approach for either quantitative or qualitative traits. The proportion of alleles shared IBD ($\hat{\pi}$) by the sibs is estimated from the information at the marker locus. This proportion can be 0, ½, or 1 which corresponds to sharing 0, 1, or 2 alleles IBD. This is regressed on the squared difference in the phenotype at the trait locus. If linkage is present between the marker and the trait, then the phenotypes of the sibs are more similar if they share more alleles IBD; therefore, the difference tends to be smaller when the sibs share two alleles and larger when they share no alleles. To test the hypothesis of no linkage, we simply test whether the regression coefficient is different from 0.

Sib-pair methods have been used to implicate several chromosomal regions in genetic mapping studies of complex traits. One example is a region on chromosome 2 that may contain a gene for non-insulin dependent diabetes mellitus (NIDDM). This region was identified using the ASP analysis with 330 sib-pairs (Hanis et al., 1996); however, this region has not been confirmed in other data sets. In addition, at least 5 regions have been suggested to play a role in insulin dependent diabetes mellitus (IDDM) using the ASP analysis (Davies et al., 1994; Field et al., 1994; Hashimoto et al., 1994; Luo et al., 1995; Zamani et al., 1996; Copeman et al., 1995). None of the regions were confirmed by every study, but several regions were detected in more than one study.

There are some disadvantages to sib-pair approaches as well. One problem with complex traits is that they are often adult-onset diseases, which can increase the amount of missing data, especially in the older generations. This makes it difficult, if not impossible, to use methods that require knowledge of the IBD status. If the parental genotypes are unknown, then we often only have identity by state (IBS) information rather than the IBD status for the sib-pairs. Bishop and Williamson (1990) showed that the power of IBS information is much lower than the power of IBD status to detect linkage. In addition, sib-

pair methods are often less powerful than the lod score method, particularly if the model parameters can be correctly specified (Goldin and Weeks, 1993). Therefore, it may not be possible to detect some loci that contribute to a complex trait with a reasonable sample size using sib-pair methods.

1.3.4 Extensions of sib-pair methods

For quantitative traits, Carey and Williamson (1991) proposed a sampling method called the single proband sib-pair (SPSP) approach to increase the power compared to random sampling of sib-pairs. In this method, the sample ascertainment proceeds by selecting one sib with a high trait value (assuming a high trait value corresponds to being "affected"), and randomly selecting the other sib. This sampling method is more powerful than unselected sib-pairs for all of the recessive and additive models tested. For the dominant models, the SPSP approach is more powerful if the allele frequency is low (less than 0.6).

Risch and Zhang (1995) proposed a different sampling method called the double proband sib-pair approach (DPSP) to increase the power of the test for quantitative traits. They showed that the sample size necessary to detect linkage can be greatly reduced if one uses pairs that are either extremely discordant (ED) or extremely concordant (EC) in phenotype. Sibs are extremely discordant if one sib has a high value of the quantitative trait (e.g. in the top 10%), while the other sib has a low value (e.g. in the bottom 10%). Sibs are extremely concordant when both sibs have high values or both sibs have low values of the quantitative trait. Using the traditional sib-pair approach, a locus must account for >50% of the variance to be detectable with reasonable sample sizes (Blackwelder and Elston, 1982). However, with the DPSP approach, it may be possible to detect loci that account for as little as 10% of the variance using reasonable sample sizes. The disadvantage to the DPSP approach is that the number of sib-pairs that must be screened in order to obtain the sample of ED pairs might be very large (in the thousands) depending on the size of the genetic effect (Risch and Zhang, 1996).

1.3.5 ASSOCIATION STUDIES

Association studies have been proposed as a possible alternative to family based studies (Risch and Merikangas, 1996; Brown and Hartwell, 1998). There are several disadvantages to the family based approach in human populations. It is often difficult to collect informative families because the family members are unwilling to participate in the study. In addition, we must select matings from those that exist in the population, which may not be completely informative. Finally, missing information or incomplete information in the pedigree can significantly increase the computational burden of the analysis. In association studies, we avoid these problems by collecting independent cases and controls, instead of requiring related individuals.

Linkage disequilibrium is a population-based association between allele(s) at a marker locus and allele(s) at a trait locus. This association may be caused by the close proximity of the loci to each other, or linkage. If two loci are linked, they are usually transmitted together except for a small part of the time that depends on the distance between the loci. As time increases, this association is disrupted, so the mutation at the trait locus must also be relatively recent in order for linkage disequilibrium to be detectable. However, an association between two loci can also be caused by methods other than linkage such as admixture or population stratification (Wijsman, 1997).

There are several disadvantages to association studies. In order to determine if the association is from linkage rather than another cause, we must be able to evaluate whether the marker alleles are transmitted with the disease phenotype. In association studies that use independent individuals, it is not possible to evaluate the transmission of alleles. In addition, it is not certain that linkage disequilibrium will be present or detectable for every disease. This is particularly true for complex traits where there may be multiple causes of the disease. Even if there is linkage disequilibrium present for one form of the disease, it may not be detectable in a mixed sample of individuals. Finally, a large number of mark-

ers must be evaluated in this method because the region surrounding the trait locus where linkage disequilibrium is present and detectable may be very small.

1.4 STUDY DESIGN ISSUES

Within the context of complex traits, it is unclear which strategies are the most effective for identifying genes that contribute to the development of disease. In order to compare strategies, we must have some criteria to judge which method is better in a particular situation. Usually, power and sample size considerations are used as the main criteria for comparing methods. This criterion may grossly underestimate the utility of some designs. For example, sib-pair methods often require a larger sample size to detect linkage than lod score methods; however, sib-pairs may be significantly easier to identify and to collect than the more extended pedigrees that are often used with lod score methods.

Additional factors can be incorporated into the evaluation of study designs by considering the overall cost of each design. This strategy has been used recently by several other investigators to explore the cost-effectiveness of alternative study designs (e.g. Elston et al., 1996; Gu et al., 1996). The results of these studies demonstrate that it is important to incorporate these additional cost factors into the evaluation of the relative performance of alternative study designs. There are several factors that contribute to the cost of a study that we may want to consider including the cost of sample collection (including phenotyping), the cost of marker genotyping, and the cost of the statistical analysis.

We evaluate each of these factors to identify cost-effective strategies for the analysis of complex traits. The overall cost is used here as the criteria for evaluating alternative study designs for a given power and significance level. The particular issues that are addressed in this dissertation are described here.

1.4.1 The effect of family configuration on the cost of linkage analysis

Pedigree selection can have a large impact on the overall cost of a study. Complex traits are often difficult to study because of the heterogeneity present in the sample. We demonstrate that the number of affected and unaffected individuals observed in the pedigree can be used to identify which pedigrees are more likely to be segregating a particular locus, even among pedigrees of the same size and structure. Therefore, the ascertainment scheme used to collect the sample can result in a more homogeneous sample, with greater resulting power to detect linkage. We provide an analytical framework for evaluating the relative utility of different nuclear pedigrees. In addition, we present results of a simulation study comparing selection schemes for an extended pedigree. Although the optimal pedigree structures that provide the most information about linkage at a particular locus are highly model dependent, we identify ascertainment schemes that are cost-effective for a wide variety of underlying models. These results are presented in Chapter 2.

1.4.2 Characteristics of a genetic map for a cost-effective genome screen using diallelic markers

Many new technologies for genotyping SNP markers have recently been developed that may eventually be lower in cost, and more easily automated than current methods. Studies of complex traits require large sample sizes, so a reduction in the cost of genotyping may have a significant impact on the overall cost of a study. Using a combination of analytic and simulation methods, we explore characteristics of uniform vs. clustered SNP marker maps, and evaluate the cost of a study relative to the cost using STRP marker maps. Issues that are addressed in comparing the map structures include the information content for clustered or single markers, and the map accuracy. These results are presented in chapter 3.

1.4.3 Reduction in computation time from downcoding markers

The computation time necessary to perform a statistical analysis often limits the analyses that can be performed on a given sample. Studies of complex traits often have many characteristics that contribute to a large computational burden including missing data in the pedigrees, model complexity, and large data sets. One method that can be used to reduce the computational burden is marker allele reduction, or downcoding. Markers with a large number of alleles are desirable for genetic studies since they are generally more informative for linkage; however, in many cases, not all of the observed alleles are necessary to provide the maximum amount of information available in the pedigree. Downcoding is often required for powerful linkage analysis methods such as two-point or multipoint linkage analysis, and may have a significant impact on the cost of the analysis compared to alternative methods. In addition, downcoding may allow the use of analysis methods in situations that were previously impossible within the constraints of the computer resources available. In chapter 4, we describe a method of downcoding that can be used in conjunction with many of the statistical analyses that are currently available.

CHAPTER 2 : THE EFFECT OF FAMILY CONFIGURATION ON THE COST OF LINKAGE ANALYSIS

2.1 INTRODUCTION

Many complex traits have multiple genetic and non-genetic causes. For many diseases, it may not be possible to distinguish the different forms of the disease on the basis of the phenotype. Therefore, a given sample may include multiple genetic, as well as sporadic forms of the disease. The heterogeneity inherent in such a sample creates problems for the genetic mapping and cloning studies that are becoming increasingly popular as a way of identifying the underlying causes of complex diseases. Numerous studies have shown that the power to detect linkage to a particular trait locus is substantially decreased when there is heterogeneity both within and between pedigrees in the sample regardless of the analysis method that is used (*e.g.* Risch, 1990; Goldin and Weeks, 1993; Goldin and Gershon, 1988). A reduction in power results in an increase in the sample size necessary to detect linkage, which increases the cost of the study. In addition, the estimate of the disease location obtained with model-based methods may be biased if the sample is heterogeneous, leading to difficulties in cloning the relevant genes.

A number of different strategies have been used to try to account for this heterogeneity. One method is to select large pedigrees with multiple affected individuals to attempt to enrich for genetic forms compared to sporadic forms of the disease. This strategy has been employed for diseases such as early-onset Alzheimer's disease (Schellenberg *et al.*, 1993), and prostate cancer (Carter *et al.*, 1992; Eeles *et al.*, 1998). Sometimes only a single large pedigree is collected to increase the probability that there is a single cause of the disease in the sample (Vahava *et al.*, 1998). At the opposite end of the spectrum, study designs based on affected sib-pair (ASP) methods are often used for the analysis of complex traits. Misspecification of model parameters for parametric linkage analysis can lead

to a loss of power to detect linkage (Clerget-Darpoux *et al.*, 1986; Vieland *et al.*, 1992a,b; Greenberg and Hodge, 1989). For complex traits it is often difficult to correctly specify the model parameters. However, ASP methods do not require the estimation of many of the genetic parameters in order to carry out the linkage analysis. In addition, ASPs are often easier and less expensive per individual to collect than are more extended pedigrees. Examples of complex traits that have been studied using ASP methods include diabetes (Davies *et al.*, 1994; Imperatore *et al.*, 1998), bipolar affective disorder (Stine *et al.*, 1997), and schizophrenia (Daniels *et al.*, 1997; Garner *et al.*, 1996; Pulver *et al.*, 1995).

The large pedigree sampling strategy described above implies that different sampling strategies may result in different mixtures of the genetic and non-genetic forms of the disease in the sample. This suggests that one solution to the problem of sample heterogeneity in complex traits may be to create more homogeneous samples using pedigree selection criteria. The issue of strategies to reduce sample heterogeneity is almost never considered in study designs using ASP methods. This is unfortunate since these methods are particularly useful for complex traits where there is almost certainly heterogeneity present in the sample. Factors such as the distribution of affected status among the offspring and the pedigree size could be used as criteria for pedigree selection. Several studies have investigated the power of different sampling schemes within the context of nuclear pedigree structures (McCarthy *et al.*, 1998; Goldgar and Easton, 1997; Ginsburg and Axenovich, 1996; Sribney and Swift, 1992; Durner *et al.*, 1992). However, these studies have been limited in the pedigree structures and/or the sampling schemes that were considered.

The choice of sampling schemes should depend not only on power and sample size considerations, but also on the cost of obtaining and genotyping the sample. A smaller sample size will not necessarily result in a lower cost. For example, although large pedigrees are often more informative for linkage than nuclear pedigrees, selection schemes that require larger pedigrees may still be more expensive if large pedigrees are less common

and more difficult to collect. There have been several other studies that have incorporated cost into the evaluation of study designs (Elston *et al.*, 1996; Goldgar and Easton, 1997; Gu *et al.*, 1996). These studies demonstrate that study design comparisons are affected by these additional cost factors.

In this chapter, we address two components pertinent to pedigree selection for gene mapping studies of complex traits. Our first goal is to demonstrate that pedigrees with different numbers of affected and unaffected individuals differ in the amount of information they provide for linkage analysis at a particular trait locus, even if only a single ASP per pedigree is used in the analysis. First, we focus on ASP designs where we show that different sampling schemes alter the fraction of ASPs segregating for genes at different trait loci. This results in differing power to detect linkage among sampling schemes, and may have implications for the failure to replicate linkage results in different studies. We provide an analytical framework for evaluating the difference in power among selection schemes. We also present results from a simulation study to evaluate the effect of sampling schemes on the power to detect linkage in a large extended pedigree. Our second goal is to demonstrate that these results can be used to identify selection schemes that lower the cost of a study for a wide variety of underlying genetic models. This is important since generally we do not know the true underlying model ahead of time, particularly for complex traits. The present study extends the results of previous authors by considering a broader range of possible pedigree structures, as well as sampling schemes that depend on the number of affected and unaffected individuals in the pedigree. In addition, we consider the overall cost of the study instead of comparing study designs based solely on the sample size requirements.

2.2 METHODS

Our first goal in this analysis is to demonstrate that alternative pedigrees differ in the amount of information they provide for linkage analysis within the context of ASP meth-

ods. The means test is the ASP approach that will be used, since this test is more powerful than many other ASP methods (Blackwelder and Elston, 1985; Davis and Weeks, 1997). The power of this approach depends on the expected proportion of alleles shared IBD (Y) under the alternative hypothesis of linkage. Large deviations from the probability expected under the null hypothesis (0.5) result in a high power, and a small sample size to detect linkage. We show that the expected proportion of alleles shared IBD is not the same for all ASPs under a given model, but depends on the number of siblings with known phenotype in the pedigree (n), and the number of affected individuals (a) in the sibship. Here, we define the term *family configuration* to refer to the sibship size, and the distribution of affected and unaffected individuals in the sibship. Note that the phenotype of the unaffected individuals is used to determine the family configuration, but only the genotypes of the affected sibs and the parents are used in the analysis.

Suarez *et al.* (1978) first calculated the expected proportion of alleles shared IBD for a random ASP when the phenotypes of relatives other than the sib-pair are not considered. The expected proportion of alleles shared IBD has also been calculated for the single-locus case (Sham *et al.*, 1997) and for the multi-locus case (McCarthy *et al.*, 1998) conditional on the family configuration. We present results that demonstrate how the relationship between the expected proportion of alleles shared IBD and the family configuration depends on different parameters of the underlying genetic model.

2.2.1 Expected proportion of alleles shared IBD by an ASP

The expected proportion of alleles shared IBD (Y_j) at trait locus j for a randomly chosen ASP is calculated conditional on the total number of siblings with known phenotype in the pedigree, and the number of affected individuals in the sibship. The parental phenotypes are not considered here in determining Y_j, since this information is often not available for many traits. If there are multiple loci involved in determining the trait phenotype, we assume the disease loci are unlinked, and there is no linkage disequilibrium between the loci. Note that we are calculating Y_j at the trait locus. In most studies, we are not able

to measure the trait locus, and we would need to consider the effects of recombination between a marker locus and the trait locus, and incomplete information for the marker.

Define z_{ij} to be the probability that a randomly chosen ASP from a given family configuration shares i alleles IBD for $i = 0$, 1, or 2 at trait locus j:

$$z_{ij} = P(\pi_j = i \mid n,a).$$

Here, π_j is the number of alleles shared IBD at trait locus j by the ASP, and the family configuration is defined by n and a. The z_{ij} is calculated by summing over the probability for each mating type (M), where the mating type is the multi-locus genotypes for the parents:

$$z_{ij} = \frac{\sum_M P(n,a|\pi_j = i, M)P(M)P(\pi_j = i)}{\sum_M P(n,a|M)P(M)}.$$

The term $P(n,a|\pi_j = i,M)$ is calculated by summing over all possible pairs of genotypes for the two affected individuals that share i alleles IBD. For the remaining n-2 individuals, we can determine the probability that there are a-2 affected individuals and n-a unaffected individuals given the probability that an offspring is affected for the M mating type. When there are multiple loci in the model, z_{ij} is calculated separately for each locus, where the value of π_j is not specified for the other loci in the model. Y_j can then be calculated using the values of z_{0j} and z_{1j}:

$$Y_j = 1 - z_{0j} - \tfrac{1}{2} z_{1j}.$$

For the remainder of the chapter, we do not include the subscript j unless there are multiple loci in the model.

2.2.2 Calculation of cost

Our second goal is to demonstrate that the information on the relationship between the expected proportion of alleles shared IBD and the family configuration can be used to identify selection schemes that lower the cost of a study for a wide variety of underlying genetic models. We present a method of combining the costs of pedigree collection, phenotyping, and genotyping into an overall cost for each ascertainment scheme. Finally, we describe four ascertainment schemes that are compared here using this cost function.

2.2.2.1 Probability of nuclear family configuration

The probability of observing a particular nuclear family configuration depends on the distribution of sibship size, and the underlying genetic model. The distribution of sibship size is generally not known, and is also not constant between populations, or generations within a population. We chose to model the distribution of sibship size using a truncated geometric distribution ($n < 11$) with parameter d as suggested by Vogler *et al.* (1995) despite the fact that this model may not be appropriate for every population. In this model, the probability that the sibship has x individuals is:

$$P(n = x) = \frac{(1-d)^{x-1} d}{\sum_{i=0}^{10} (1-d)^{x-1} d}$$

Most of the results presented here are for a model with $d = 0.45$; however, we also considered a model where $d = 0.2$ to determine the effect of a larger overall sibship distribution.

The probability of observing a out of n affected offspring can be calculated by using the binomial distribution, given the probability of being affected for each mating type (p_M). The probability of observing a particular nuclear family configuration can then be found by summing over all possible mating types (M):

$$P(n,a) = P(a|n)P(n) = \sum_M P(a|n,M)P(M)P(n) = \sum_M \binom{n}{a} p_M^a (1-p_M)^{n-a} P(M)P(n).$$

2.2.2.2 Number of independent pairs

Once we have calculated the expected proportion of alleles shared IBD, we can then calculate the sample size necessary for a given power and significance level for any test statistic based on the expected proportion of alleles shared IBD. Although many tests have been proposed in the context of the ASP design, the only test considered here is the means test. We can define a variable t_i that equals 0, 1, or 2 depending on the number of alleles shared IBD by the ith sib-pair. The expected value of t_i under the null hypothesis is 1, and the variance is ½. We can construct the test statistic, T, which has an asymptotic standard normal distribution

$$T = \frac{\sum t_i / n - 1}{\sqrt{1/2n}}.$$

Under the alternative hypothesis, the expected value and the variance (σ^2) of T are $\sqrt{2n}(2Y-1)$ and $4z_2 + 4Y(1-2Y)$ respectively (see appendix). The variance was calculated assuming the t_i's are independent. This assumption is violated if all possible ASPs per sibship are used when there are more than two affected individuals per sibship. We do not account for the dependence between ASPs in this case, so the sample size may be underestimated. However, the sample size required when only the independent ASPs per sibship are used will provide an upper bound on the sample size when all possible ASPs per sibship are used. For other methods of including multiple ASPs per sibship, this assumption may be valid. The normal approximation was used to determine the number of independent sib-pairs (N) needed for a power of $1-\beta$ and a significance level of α:

$$N = \frac{(Z_\alpha - \sigma Z_{1-\beta})^2}{2(2Y-1)^2}.$$

For the examples presented here, we assumed $Z_\alpha = 3.72$ (α=0.001) and $Z_{1-\beta} = -0.84$ (1-β = 0.8). As we mentioned previously, we are calculating Y at the trait locus, in other words, we are assuming a completely informative marker with no recombination. If we consider the effects of recombination and reduced marker information, the necessary sample size increases.

For most ascertainment schemes, there are multiple family configurations represented in the sample. We can calculate $\overline{Y}$ for the sample of family configurations by taking a weighted average of Y for each family configuration. The weights depend on the probability of observing each family configuration (P(n,a)), and the number of affected sib-pairs that are available from each family configuration. The weights are equal to

$$\frac{aP(n,a)w(a)}{\sum_n \sum_a aP(n,a)w(a)},$$

where $w(a)$ is the number of ASPs per pedigree with a affected individuals for a given method of including multiple ASPs per sibship (e.g. for the ALL method $w(a) = a(a-1)/2$), and the sums in the denominator are for all values of n and a that are possible for a given ascertainment scheme. Similarly, we can also find a weighted average for the value of z_2. The expected value and the variance of the test statistic under the alternative hypothesis depend on the observed distribution of family configurations, so the required sample size also depends on the observed distribution of family configurations. We calculate the sample size under the expected distribution of family configurations, so the averaged values of $\overline{Y}$ and $\bar{z}_2$ (for σ) are used to calculate N for a given ascertainment scheme. The exact sample size and cost will differ depending on the actual observed family configurations in the sample. The sample size using the normal approximation described here is similar to the sample size found using an exact computation for the models evaluated (Chapman and Wijsman, 1998).

2.2.2.3 Allowing for Multiple Affected Individuals per Sibship

When there are more than two affected individuals per sibship there are multiple, dependent ASPs that can be used in the analysis. For a pedigree with a affected individuals there are $a(a\text{-}1)/2$ possible ASPs, and a-1 independent ASPs. We determine the number of families necessary to obtain N ASPs for analysis based on three methods: one ASP per sibship (EQUAL), all possible ASPs per sibship (ALL), and all independent ASPs per sibship (IND). Here we are referring to the number of ASPs per sibship used in the analysis. There may be additional ASPs present in the sibship that are not used in the analysis. There are many other possible methods for including multiple ASPs per sibship (Hodge, 1984; Suarez and Van Eerdewegh, 1984; Sham *et al.*, 1997); however, the three methods above span the possible range from only one ASP to all possible ASPs. Since there does not seem to be a general consensus on the best way to include multiple ASPs per sibship, we will evaluate all three methods.

We can calculate the expected number of each family configuration ($N_{n,a}$) that is observed in the sample for a given ascertainment scheme given the probability of each family configuration ($P(n,a)$), the number of independent ASPs (N), and a particular method of including multiple ASPs per sibship. For the ascertainment schemes that are considered here, we assume that we have a complete list of all affected individuals in the population, such as might exist in a complete disease registry. We assume the pedigrees are sampled through probands in this list. The probability of sampling a particular family configuration is therefore proportional to the probability of sampling a proband from that family configuration which depends on the number of affected individuals, and the probability of observing the family configuration in the population. Therefore, $N_{n,a}/(aP(n,a))$ is a constant for all possible pairs of values for n and a under the given ascertainment scheme. Also, if w_a is the number of ASPs that are included per pedigree with a affected individuals, then, $\sum_n \sum_a w_a N_{n,a} = N$. We can solve this series of equations to calculate the expected number of each family configuration, $N_{n,a}$.

2.2.2.4 Cost equation

The cost associated with a particular ascertainment scheme depends on the cost associated with phenotyping and collecting the individuals, and the cost of genotyping the individuals. We can write the overall cost of the study (C) as:

$$C = \Sigma_n \Sigma_a N_{n,a}\{f(n)C_P + f(a)C_G\},$$

where C_P is the average cost of phenotyping and collecting an individual including the cost of ascertaining the proband, and C_G is the average cost of genotyping an individual. We assume that the cost of phenotyping the pedigree will depend on the family size through some function, $f(n)$. The cost of genotyping depends on the number of individuals that must be genotyped. For example, in the EQUAL method only two affected siblings and the parents are genotyped per pedigree so $f(a) = 4$, while for the IND or ALL methods all affected siblings and the parents are genotyped per pedigree so $f(a) = a + 2$. If the parents are not genotyped, we potentially lose some information about the IBD status of ASPs. This loss of information would not be uniform across all family configurations, so we might need to consider this in the cost analysis if we did not assume that the parents are genotyped. For the two-locus models, we are considering the cost of mapping the *first* locus thus mapped, so we include the cost of an entire genome scan, even though the marker information could be used later to identify additional loci. We divide the total cost (C) by the cost of genotyping an individual (C_G) to find the cost relative to the cost of genotyping (C^*) :

$$C^* = \frac{C}{C_G} = \Sigma_n \Sigma_a N_{n,a}\{f(n)\frac{C_P}{C_G} + f(a)\} = \Sigma_n \Sigma_a N_{n,a}\{f(n)R + f(a)\},$$

so R is the relative cost of collecting and phenotyping an individual compared to the cost of genotyping an individual.

The values of R that we considered (0.1, 1.0, and 10) correspond to a range of values that might be observed in genetic studies. One of us (EMW) estimated R for two studies, one on dyslexia, and one on Alzheimer's disease. For the study on dyslexia, the estimated cost of genotyping for a 10 cM genome screen is $650 per person, while the estimated cost of phenotyping and collecting the individuals is $900 per person. Therefore, a value of $R \cong 1.0$ is appropriate. For the study on Alzheimer's disease, the estimated cost of phenotyping and collecting the individuals is $3300 - $5000 per person. In this case, the value of R is closer to 10. Alternatively, it has been argued that some phenotypes may be much easier to measure than these complex phenotypes (*e.g.* blood pressure, weight), in which case a value of $R \cong 0.1$ is more appropriate.

We evaluate three functions to describe the effect of family size on the cost of ascertainment. First, we use $f(n) = n$ (LINEAR) for the relationship between family size and the cost of phenotyping. In this case, the cost of phenotyping is directly proportional to the family size. Alternatively, large families may be much harder to collect per person than small families. For example, not everyone in the pedigree may be interested in participating in the study, or the family members may not all be located in one region. In this case, we use $f(n) = n^2$ (SQUARE) for the relationship between the family size and the cost of phenotyping. Finally, there might be situations where large families are relatively easier to collect per person, although still more expensive in total because there are more individuals. This might occur if the families are highly motivated because of the large number of affected relatives, or if the families are ascertained through a clinic where individuals are more likely to be referred if they have other affected family members. In this case, we use $f(n) = \sqrt{n}$ (SQRT) for the relationship between the family size and the cost of ascertainment.

2.2.2.5 Ascertainment schemes

We consider four ascertainment schemes that depend on the number of affected and unaffected individuals in the sibship. We assume that the probability that a pedigree is sam-

pled is proportional to the probability of observing the pedigree in the population. Pedigrees are selected if they meet the following criteria:

$S_{2,0}$. Pedigrees with at least 2 affected individuals.
$S_{2,1}$. Pedigrees with at least 2 affected individuals, and at least 1 unaffected individual.
$S_{3,0}$. Pedigrees with at least 3 affected individuals.
$S_{3,1}$. Pedigrees with at least 3 affected individuals, and at least 1 unaffected individual.
Scheme $S_{2,0}$ is a common selection scheme that is used for collecting samples of ASPs, where all that is required is that there is a single ASP in the pedigree. We chose the remaining three schemes to evaluate the effect of selecting for additional affected and/or unaffected individuals on the power to detect linkage.

2.2.3 Model parameterization

2.2.3.1 Single-locus models

We consider a single diallelic trait locus with alleles D and d, and allele frequencies p_D and p_d respectively. The genotype frequencies for genotypes DD, Dd, and dd are denoted g_{DD}, g_{Dd} or g_{dd}, and the penetrances are denoted f_{DD}, f_{Dd}, and f_{dd} respectively. The set Ω contains the genotypes that are associated with the genetic form of the disease (*e.g.* for a dominant model Ω = {DD, Dd}, and for a recessive model Ω = {DD}). $\bar{\Omega}$ is the complement of Ω. The sporadic proportion (s) is defined as:

$$s = \frac{\sum_{i \in \bar{\Omega}} f_i g_i}{\sum_{i \in (\Omega \cup \bar{\Omega})} f_i g_i},$$

where the sporadic proportion is the probability that an individual is affected due to some cause of disease other than a mutation at the genetic locus in the model. The source of the sporadic cases remains unspecified but includes other unknown loci, environmental factors, or misclassification of individuals.

2.2.3.2 Two-locus epistatic model

For the epistatic model there are two diallelic trait loci, locus A with alleles 'A' and 'a', and locus B with alleles 'B' and 'b', where p_A, p_a , p_B, and p_b are the respective allele frequencies. The two loci are assumed to be unlinked and in linkage equilibrium, so the probability of the joint genotype is the product of the genotype probabilities for the individual loci. The penetrances, f_{ij}, depend on the joint genotype where $i \in \{AA, Aa, aa\}$ and $j \in \{BB, Bb, bb\}$. The penetrances for the genetic form of the disease are the product of the marginal penetrances such that $f_{ij} = f_i \cdot f_j$. Similar to the single locus model, there may also be a sporadic proportion of disease for this model. The sporadic proportion is defined as above for the single locus model.

2.2.3.3 Two-locus heterogeneity model

For this model, the alleles and the allele frequencies are defined as above for the epistatic model. Again, we assume that the trait loci are unlinked and in linkage equilibrium so the probability of the joint genotype is the product of the probabilities of the single locus genotypes. In the heterogeneity model, the loci act independently to cause disease, so a mutation at only one of the loci is required for the disease. The penetrance of the joint genotype is $f_{ij} = f_i + f_j - f_i f_j$, where $i \in \{AA, Aa, aa\}$ and $j \in \{BB, Bb, bb\}$. The sporadic proportion is defined as above for the single locus model.

2.2.4 Simulation study for extended pedigree analysis

A simulation study was used to evaluate the relative costs of different ascertainment schemes for an extended pedigree. A fixed pedigree structure was used for this analysis that was composed of 16 individuals in three generations. There were four offspring in the middle generation, and 4 individuals in each of two sibships in the bottom generation. The ascertainment criteria were based on the number of affected individuals in the pedigree. A pedigree was selected for inclusion in the analysis if there were $\geq i$ affected indi-

viduals in the pedigree for $i = 1, \ldots, 5$. The affected individuals could be anyone in the pedigree.

Ten two-locus heterogeneity models were evaluated. The model parameters are given in Table 2.1. The models were selected to have a population prevalence between 1-10%, and a relative risk to sibs between 2 and 5. We considered these parameter values to be representative of common, complex diseases. All of the models had a dominant mode of inheritance at both trait loci, but the relative penetrance and the allele frequencies varied among the models. Trait phenotypes and genotypes were simulated for the fixed pedigree structure until 500 pedigrees meeting the selection criteria were found. The affected individuals in the simulated sample were then classified into the following classes: 'A only' for individuals with genotypes AAbb and Aabb, 'B only' for individuals with genotypes aaBB and aaBb, 'A and B' for individuals with genotypes AABB, AABb, AaBB, and AaBb, and 'sporadic' for individuals with genotype aabb.

Single-trait locus parametric linkage analysis was used as the analysis method since a sib-pair analysis would not be efficient for the extended pedigree structure. PAP (Hasstedt and Cartwright, 1981) was used to obtain estimates of the model parameters for the analysis. The lod scores were summed over n families where n was allowed to vary. To determine the necessary sample size we found the minimum value of n such that the sum of the lod scores for at least 80% of the n families exceeded some threshold. Pedigrees with more affected individuals may be more difficult to recruit because they are not found as often in the population. Therefore, we assumed the cost of recruiting a pedigree with a affected individuals is inversely proportional to the frequency of pedigrees with a affected individuals in the population for the cost analysis. The total cost (C) depends on the number of families with each number of affected individuals (N_a), the cost of recruitment (R), the cost of phenotyping (P), and the cost of genotyping (G). The costs of phenotyping and genotyping are fixed costs for the given pedigree structure, and we can

write these costs as a multiple of the cost of recruitment (kR). Therefore, the overall cost is $C = \sum_s N_s(R+P+G) = \sum_s N_s R(1+k)$.

2.3 RESULTS

2.3.1 EXPECTED PROPORTION OF ALLELES SHARED IBD

2.3.1.1 Single-locus models

We evaluated a total of 125 single locus models to determine the effects of penetrance, allele frequency, mode of inheritance, and sporadic proportion on the relationship between family configuration and the expected proportion of alleles shared IBD. The penetrance ranged from 0.025 to 0.9, the allele frequency ranged from 0.01 to 0.04 for dominant models and from 0.025 to 0.3 for recessive models, and the sporadic proportion ranged from 0 to 0.9. Dominant, recessive, and codominant modes of inheritance were all evaluated. The general results are consistent across the models evaluated, so only a few examples are presented here to demonstrate the main observations.

The sporadic proportion has the largest effect on the relationship between family configuration and the expected proportion of alleles shared IBD relative to the effects of the other parameters evaluated. When s = 0, family configurations with a low number or a low density (*a/n*) of affected individuals have the highest expected proportion of alleles shared IBD (Figure 2.1A,D). For example, in figure 2.1A, pedigrees with 2 out of 6 siblings affected have Y equal to 0.75, while pedigrees with 6 out of 6 siblings affected have a lower value of Y equal to 0.68. Parents in family configurations with fewer affected individuals or a low density of affected individuals are more likely to be heterozygous for the high-risk allele, so the expected sharing among the ASPs is higher than for other family configurations. As s increases, family configurations with a high number or a high density of affected individuals have the highest expected proportion of alleles shared IBD (Figure 2.1B,C,E,F). For example, in figure 2.1C, pedigrees with 2 out of 6 siblings af-

fected have Y equal to 0.64, while pedigrees with 6 out of 6 siblings affected have a higher value of Y equal to 0.67. The cases in these family configurations are less likely to be sporadic cases, so the expected proportion of alleles shared IBD is higher than in family configurations with only a few affected individuals. Not surprisingly, the expected proportion of alleles shared IBD is lower for all family configurations as the sporadic proportion increases. At least some of the affected sibs in many families are sporadic cases when s is high, and such sibs will not share alleles IBD with other affected sibs more often than expected by chance.

The penetrance also has an effect on the relationship between the expected proportion of alleles shared IBD and family configuration. For a given number of affected sibs, the expected proportion of alleles shared IBD increases more quickly with additional unaffected sibs when the penetrance is high (Figure 2.1B *vs.* 2.1E), since unaffected sibs provide more information about the parental genotypes in this case. The addition of unaffected sibs to the pedigree increases the probability that the parents are heterozygous, which also increases the expected proportion of alleles shared IBD for the ASPs. For example, in figure 2.1A (high penetrance model), if all 6 siblings in a pedigree are affected Y equals 0.677, while the addition of one unaffected sibling increases Y to 0.725. However, in figure 2.1D (low penetrance model), Y equals 0.677 if all 6 siblings are affected, but Y only increases to 0.686 with the addition of one unaffected sibling.

The remaining two parameters, allele frequency and mode of inheritance, do not change the relationship between family configuration and the expected proportion of alleles shared IBD; however, these parameters do cause a shift in the expected proportion of alleles shared IBD for a given family configuration. As might be expected, high allele frequencies generally reduce the expected proportion of alleles shared IBD. For example, in figure 2.2B (lower allele frequency), the maximum value of Y is 0.724, while in figure 2.2E (higher allele frequency), the maximum value of Y is 0.699. Multiple copies of the disease causing allele are more likely to be segregating in the pedigree when the allele frequency is high which tends to reduce the expected proportion of alleles shared IBD

among ASPs. Out of the modes of inheritance evaluated, the recessive model has the highest expected proportion of alleles shared IBD (Figure 2.3). For a recessive disease, affected individuals must have two copies of the high-risk allele, whereas for a dominant or codominant disease only one copy is required. These results are consistent with observations by others on the relative power of the ASP approach for recessive *vs.* dominant models (Goldin and Gershon, 1988; Goldin and Weeks, 1993).

2.3.1.2 Two-locus epistatic models

We evaluated twenty-seven two-locus epistatic models to explore the relationship between family configuration and the expected proportion of alleles shared IBD. The effects of sporadic proportion, allele frequency, penetrance, and mode of inheritance are the same for the epistatic models as for the single-locus models. For example, Figure 2.4 depicts two epistatic models with the corresponding single-locus marginal models. In general, the single-locus models have a slightly lower expected proportion of alleles shared IBD than the epistatic models, but the overall relationship between family configuration and the expected proportion of alleles shared IBD is the same.

2.3.1.3 Two-locus heterogeneity models

For the two-locus heterogeneity models, the disease model at both loci must be considered to predict the relationship between family configuration and the expected proportion of alleles shared IBD. Three hundred and twenty-two two-locus heterogeneity models were evaluated in this analysis. Both dominant and recessive modes of inheritance were evaluated, with a penetrance between 0.1 and 0.9, and a sporadic proportion between 0 and 0.9. The allele frequency at loci with a dominant mode of inheritance was either 0.01 or 0.02, and the allele frequency at loci with a recessive mode of inheritance was between 0.01 and 0.2.

In order to describe the relationship between family configuration and the expected proportion of alleles shared IBD we must consider the relative population prevalence of the

two genetic forms of the disease. The population prevalence depends on several parameters in the model including the penetrance, the allele frequency, and the mode of inheritance. Unless the disease model is the same for both loci, there is always one locus with a higher disease prevalence; therefore, we will refer to the two loci in the model as the "higher disease prevalence" locus and the "lower disease prevalence" locus. In the example in Figures 2.5 and 2.6, the allele frequency and the mode of inheritance are the same for both loci, so the penetrance determines the relative prevalence of the two loci. The locus with the lower disease prevalence generally has a higher expected proportion of alleles shared IBD for family configurations with a low density of affected individuals, while the locus with the higher disease prevalence generally has a higher expected proportion of alleles shared IBD for family configurations with a higher number of affected individuals. For example, in figure 2.5A (lower disease prevalence locus), family configurations with 2 out of 10 siblings affected have the highest expected proportion of alleles shared IBD, while in figure 2.6A (higher disease prevalence locus), family configurations with 6 out of 8 affected individuals have the highest expected proportion of alleles shared IBD.

Once the relative prevalence of the two loci is taken into account, the effects of the other model parameters are generally the same as what we observed for the single-locus models. For example, we can examine the effect of penetrance by comparing Figure 2.5G (both loci have a low penetrance) to Figure 2.6A (both loci have a high penetrance). For a given number of affected individuals, the change in the expected proportion of alleles shared IBD due to the addition of unaffected individuals is larger for the high penetrance model (0.659 for 6/6 affected vs. 0.699 for 6/7 affected) than the low penetrance model (0.648 for 6/6 affected vs. 0.662 for 6/7 affected).

2.3.2 RELATIVE COST

Unless indicated otherwise, all of the results presented here have R = 1.0, d = 0.45, and $f(n)$ = LINEAR.

2.3.2.1 Single-locus models

The relative cost of ascertainment schemes depends on the method that is used for including multiple ASPs per sibship. For the EQUAL method, schemes $S_{2,0}$ and $S_{2,1}$ are more cost-effective if the sporadic proportion is low (Figure 2.7A), while schemes $S_{3,0}$ and $S_{3,1}$ are more cost-effective if the sporadic proportion is high (Figure 2.7C). For example, in figure 2.7 the cost of scheme $S_{3,0}$ is 1.2 times the cost of scheme $S_{2,0}$ when the sporadic probability is low (0), while the cost of scheme $S_{3,0}$ is 0.4 times the cost of scheme $S_{2,0}$ when the sporadic probability is high (0.9). These results are predictable based on the relationship between family configuration and the expected proportion of alleles shared IBD under each model (Figure 2.1). For the ALL and IND methods, schemes $S_{3,0}$ and $S_{3,1}$ are always more cost-effective than schemes $S_{2,0}$ and $S_{2,1}$. One factor that contributes to the reduction in cost for schemes $S_{3,0}$ and $S_{3,1}$ for the ALL and IND methods is that family configurations with a large number of affected individuals are more efficient under these schemes because fewer pedigrees need to be collected to obtain a given number of ASPs.

Not surprisingly, the sporadic proportion has the largest effect on the relative cost of the ascertainment schemes. In general, there is a reduction in cost for schemes $S_{3,0}$ and $S_{3,1}$ relative to scheme $S_{2,0}$ as the sporadic proportion increases (Figure 2.7). For example, in this figure, the relative cost of scheme $S_{3,0}$ compared to scheme $S_{2,0}$ is 0.82, 0.69, and 0.31 for sporadic probabilities 0, 0.55, and 0.9 respectively using the IND method. Pedigrees with a high density of affected individuals are more informative when the sporadic proportion is high, while pedigrees with a low density or low number of affected individuals are more informative when the sporadic proportion is low (see Figure 2.1). Since pedigrees with a high density of affected individuals represent a larger proportion of the

sample for schemes $S_{3,0}$ and $S_{3,1}$, these schemes are increasingly more cost-effective than schemes $S_{2,0}$ and $S_{2,1}$ as the sporadic proportion increases.

There is little difference in cost between ascertainment schemes that differ only in the number of unaffected individuals that are required (scheme $S_{2,0}$ *vs.* $S_{2,1}$ or scheme $S_{3,0}$ *vs.* $S_{3,1}$) aside from a slight increase in the cost of scheme $S_{2,1}$ compared to Scheme $S_{2,0}$ as the sporadic probability increases. This is likely due to the fact that only a small percentage of families do not have any unaffected individuals. Therefore, ascertainment schemes that do not require any unaffected individuals result in a sample that is similar to what is obtained from schemes that require at least one unaffected individual. A similar result was obtained by McCarthy *et al.* (1998) in their comparison of sample size requirements for ascertainment schemes that depend on the presence of an unaffected parent and/or an unaffected sib compared to random ascertainment of ASPs. They found that the sample size needed is similar among these ascertainment schemes, and concluded that random ascertainment was the best strategy.

2.3.2.2 Two-locus heterogeneity models

For the two-locus heterogeneity models the relative cost of ascertainment schemes depends on the relative population prevalence of the two loci. For the locus with the higher disease prevalence (*e.g.* Figure 2.8G-I or 2.9A-C), or when the two loci have the same prevalence (Figure 2.8D-F, 2.9D-F), the effects of the model parameters are similar to what was observed for the single locus models. For the locus with the lower disease prevalence (*e.g.* Figure 2.8A-C or Figure 2.9G-I), schemes $S_{2,0}$ and $S_{2,1}$ are generally more cost-effective than schemes $S_{3,0}$ and $S_{3,1}$ if the sporadic proportion is low for all three methods of including multiple ASPs per sibship. As the sporadic proportion increases, schemes $S_{3,0}$ and $S_{3,1}$ become more cost-effective than schemes $S_{2,0}$ and $S_{2,1}$.

2.3.2.3 Effect of family size distribution

We evaluated two different models for the family size distribution, and found that there is little effect of family size distribution on the relative cost of ascertainment schemes, so the conclusions of the cost analysis are not changed by this model assumption (Figure 2.10). When the value of d is reduced, the cost of the ascertainment schemes are more similar since a larger percentage of the population is composed of pedigrees that would be selected in any of the ascertainment schemes evaluated.

2.3.2.4 Effect of cost function

We evaluated a total of nine different cost functions, and found that despite some differences, the general conclusions that are presented above do not change. The two parameters that are allowed to vary are the ratio of the cost of phenotyping and collecting an individual *vs.* genotyping an individual, $R = C_P/C_G$, and the effect of family size on the cost of ascertainment, $f(n)$. The model presented in Figure 2.11 was specifically chosen because of the relatively large differences due to the use of different cost functions, while most other models produce smaller differences between cost functions. There is virtually no effect of R for the SQRT model; however, for the LINEAR and the SQUARE models, higher values of R generally have a higher relative cost. This difference is greater for the SQUARE model than for the LINEAR model, and it is also greater for the EQUAL method compared to the ALL and IND methods. More extreme values of these two parameters may result in different conclusions; however, the values of R and $f(n)$ evaluated here span a reasonable range for each parameter. The effects of the other model parameters such as the sporadic probability do not depend on the cost function.

2.3.3 EXTENDED PEDIGREES

As with the nuclear pedigrees, the sample proportion of each form of the disease depends on model parameters such as allele frequency and penetrance. Here we evaluate the proportion of individuals with each form of the disease (A only, B only, A and B, or sporadic) instead of evaluating the expected proportion of alleles shared IBD by an ASP be-

cause there are many relationships in the pedigree other than sib-pairs. If the two loci have the same disease model (Figure 2.12C), the A only and the B only groups represent approximately the same proportion of the sample. When the two loci have the same penetrance but different allele frequencies (Figure 2.12B), the locus with the higher allele frequency (A only) represents a higher proportion of the sample than the locus with the lower allele frequency (B only). Finally, when the two loci have different penetrances (Figure 2.12A), the sample proportions depend on the ascertainment schemes. The locus with the high penetrance represents a higher proportion of the sample under ascertainment schemes that select for many affected individuals, while the locus with the low penetrance represents a higher proportion of the sample under ascertainment schemes that select for fewer affected individuals. In general, the proportion of sporadic cases in the sample decreases when pedigrees are selected for many affected individuals. Therefore, we can select for the genetic forms of the disease by selecting pedigrees with a higher density of affected individuals.

The relative cost of ascertainment schemes generally decreases as the required number of affected individuals in the selection scheme increases (Figure 2.12D-F). This is similar to our observations for the nuclear pedigrees in cases with a high sporadic proportion. One interesting difference is that for the mixed penetrance model, the relative cost is higher for ascertainment schemes with at least four affected individuals than for schemes with at least three affected individuals. A likely explanation for this observation is that for schemes with at least four affected individuals, the gain in information from requiring many affected individuals does not offset the cost of obtaining pedigrees that are less common in the population. We did not explore ascertainment schemes with at least four or more affected individuals for nuclear pedigrees, so it is unclear whether the cost increases with additional restrictions on the number of affected individuals. However, these results suggest that it is not always cost-effective to restrict the sample to very large numbers of affected individuals.

2.4 DISCUSSION

We have demonstrated that different family configurations differ in the amount of information they provide to detect linkage. For both nuclear and extended pedigrees, we show that sampling schemes that depend on the family configuration alter the proportion of individuals segregating for genes at different trait loci. This has a resulting impact on the power to detect linkage among selection schemes, which affects the required sample size and the cost of the analysis. The ability to replicate linkage results may be affected by the selection schemes used among different studies. In addition, we were able to identify selection schemes that lower the cost of a study for a wide variety of underlying models. In particular, selection schemes that require at least three affected individuals are generally more cost-effective than selection schemes that require at least two affected individuals. The exceptions are 1) studies where only one ASP per sibship is used and the sporadic probability is low, and 2) studies to identify the lower disease prevalence locus in a two-locus model. Selection schemes that differ only in the number of unaffected individuals are usually not significantly different in terms of cost. These results were obtained by considering the overall cost of a study instead of comparing study designs based solely on the sample size requirements.

We found that the relationship between family configuration and the expected proportion of alleles shared IBD is predictable from only a few model parameters. These include the sporadic proportion for all models and the population prevalence for two-locus heterogeneity models. One possibility for estimating the sporadic proportion is to use segregation analysis. Although this may not provide an accurate estimate of this parameter, it may only be necessary to know whether the sporadic proportion is "high" or "low". Parameters such as allele frequency, penetrance, and mode of inheritance do not have a large impact on this relationship. This is important since these parameters are often difficult to estimate accurately, particularly for complex models. We conclude that the optimal family configuration that provides the most information to detect linkage is not the same for all models, but depends on the model parameters.

Despite the fact that the *optimal* ascertainment scheme is model dependent, it is still possible to identify ascertainment schemes that are cost-effective for a wide variety of underlying genetic models. Among the schemes examined here, the ascertainment schemes that require at least three affected individuals are generally more cost-effective than the ascertainment schemes that require at least two affected individuals. One exception is models with a low sporadic proportion if only one ASP per sibship is included in the sample. Family configurations with a low number of affected individuals have the highest expected proportion of alleles shared IBD in models with a low sporadic proportion. These family configurations account for a larger percentage of the sample in ascertainment schemes that require fewer affected individuals. Therefore, schemes that require fewer affected individuals are more cost-effective for these models. However, if multiple ASPs per sibship are used, there is some gain in efficiency from using larger pedigrees. Therefore, schemes that require fewer affected individuals are only cost-effective if only one ASP is used per sibship. A second exception is detection of the lower disease prevalence locus in two-locus heterogeneity models. Pedigrees with a high density of affected individuals are more likely to be segregating the higher disease prevalence locus. Therefore, ascertainment schemes that require fewer affected individuals are more cost-effective for detecting the lower disease prevalence locus.

Although only a few ascertainment schemes were evaluated here, these methods can easily be extended to additional schemes given a particular cost function and ascertainment scheme of interest. For example, it might be of interest to determine whether the cost increases or decreases if we require at least four affected individuals per pedigree for the nuclear families, since we showed that for some models the cost increases among extended pedigrees. However, there are some situations where these methods would not apply. For example, Goldgar and Easton (1997) explored a study design where the high penetrance locus is identified first so that pedigrees that are segregating this form of the disease can be removed from the sample. The remaining pedigrees are enriched for the low penetrance form of the disease, so a different ascertainment scheme may be optimal.

Alternatively, we may want to identify ascertainment strategies that are cost-effective for finding both loci in the model simultaneously. It may be possible to modify the approach taken here to consider these and other strategies.

A second limitation is that we did not consider the parental phenotypes in calculating the expected proportion of alleles shared IBD, or in defining the ascertainment schemes. The difference in the expected proportion of alleles shared IBD among different family configurations can be at least partially explained by the relative probability that the parents are homozygous *vs.* heterozygous for the high-risk allele. Knowledge of the parental phenotypes may provide additional information about the parental genotypes that would allow us to more accurately calculate the expected proportion of alleles shared IBD for an ASP. However, the parental phenotypes may not be available, or provide additional information beyond what can be inferred from the phenotypes of the offspring, especially in cases where there is a complex model, when there is a large sibship, or when phenotypes can only be measured in children. Therefore, it is unclear to what extent the results will change by considering the parental phenotypes.

One implication of our results is that pedigree ascertainment schemes create yet another source of heterogeneity among studies. For many traits, several independent studies are conducted simultaneously by different investigators. These studies do not necessarily have similar pedigree sampling designs. As we have demonstrated, samples that are collected with different ascertainment strategies contain different proportions of each form of the disease. When linkage is detected in one study, failure to detect linkage in alternative samples may be a result of a difference in power to detect the original locus due to the ascertainment scheme despite a similar or larger sample size. Therefore, caution must be used in comparing results from studies that use different ascertainment schemes. For complex traits, there may be additional issues that contribute to the difficulty in confirming linkage results such as differences in the diagnostic criteria, or stochastic effects (Suarez et al., 1994). This is in contrast to simple monogenic disorders with no sporadic cases, where the proportion of selected families with the linked form of the disease is not

a function of the family configuration. In these cases we can more easily determine the power to replicate a linkage result in alternative samples.

2.5 APPENDIX

In order to calculate the variance for the test statistic T, we first need to show that the expected value of T under the alternative hypothesis (H_1) is equal to $\sqrt{2n}(2Y-1)$. To do this, we need to find the expected value of t_i (the observed number of alleles shared IBD for the ith ASP) under the alternative hypothesis. Since $z_0+z_1+z_2 = 1$, then,

$$E[t_i|H_1] = 2z_2 + z_1 = 2(1 - z_1 - z_0) + z_1 = 2\left(1 - \tfrac{1}{2}z_1 - z_0\right) = 2Y.$$

Now we can show that:

$$E[T \mid H_1] = E\left[\frac{\sum t_i/n - 1}{\sqrt{1/2n}} \middle| H_1\right] = \sqrt{2n}\left(\frac{\sum E[t_i]}{n} - 1\right) = \sqrt{2n}\left(\frac{n2Y}{n} - 1\right) = \sqrt{2n}(2Y-1).$$

Next, we show that the variance of T under the alternative hypothesis is equal to $4z_2$ + 4Y(1-2Y). First, we find the variance of t_i under the alternative hypothesis. Since

$E[t_i^2 \mid H_1] = 4z_2 + z_1 = 2z_2 + (2z_2 + z_1) = 2z_2 + 2Y$, then,

$$\text{var}[t_i \mid H_1] = E[t_i^2 \mid H_1] - (E[t_i \mid H_1])^2 = 2z_2 + 2Y - (2Y)^2 = 2z_2 + 2Y(1-2Y).$$

Using this result, we can calculate the variance for T as:

$$\text{var}[T \mid H_1] = \text{var}\left[\frac{\sum t_i/n - 1}{\sqrt{1/2n}} \mid H_1\right] = 2n\,\text{var}\left[\frac{\sum t_i}{n} - 1 \mid H_1\right] = \frac{2n}{n^2}\text{var}[\sum t_i \mid H_1] =$$

$$\frac{2n^2}{n^2}(2z_2 + 2Y(1-2Y)) = 4z_2 + 4Y(1-2Y)$$

if we assume that the t_i's are independent. This assumption is correct for the designs using one ASP per sibship, or using all independent ASPs per sibship; however, this assumption does not hold for the designs using all possible ASPs per sibship.

Table 2.1 Generating models for simulation study of extended pedigrees.

RR-sib: relative risk to the sibling of an affected individual, which is equal to the recurrence risk for a sibling divided by the population prevalence. OR-sib: odds ratio for the sibling of an affected individual, which is equal to the probability of being affected for the sibling of an affected individual divided by the probability of being affected for the sibling of an unaffected individual. α: the penetrance for individuals with an 'A' allele at locus A, β: the penetrance for individuals with a 'B' allele at locus B, γ: the penetrance for individuals with genotype 'aabb'. The penetrance for individuals with both an 'A' and a 'B' allele is the maximum of α and β.

Model	Allele Frequency		Penetrance Parameters			P(affected)	RR-sib	OR-sib
	p_A	p_B	α	β	γ			
Low Penetrance								
L1	.075	.075	.3	.3	.01	0.088	2.01	2.22
L2	.1	.01	.3	.3	.01	0.070	2.37	2.64
L3	.01	.001	.3	.3	.01	0.016	4.36	4.61
L4	.01	.01	.3	.3	.01	0.021	4.44	4.80
High Penetrance								
H1	.05	.05	.7	.7	.01	0.138	2.82	3.97
H2	.05	.01	.7	.7	.01	0.090	3.96	5.60
H3	.05	.001	.7	.7	.01	0.079	4.41	6.21
H4	.005	.005	.7	.7	.01	0.033	4.99	5.79
Mixed Penetrance								
M1	.2	.005	.9	.05	.01	0.033	4.60	5.25
M2	.1	.01	1	.3	.01	0.084	3.09	3.81

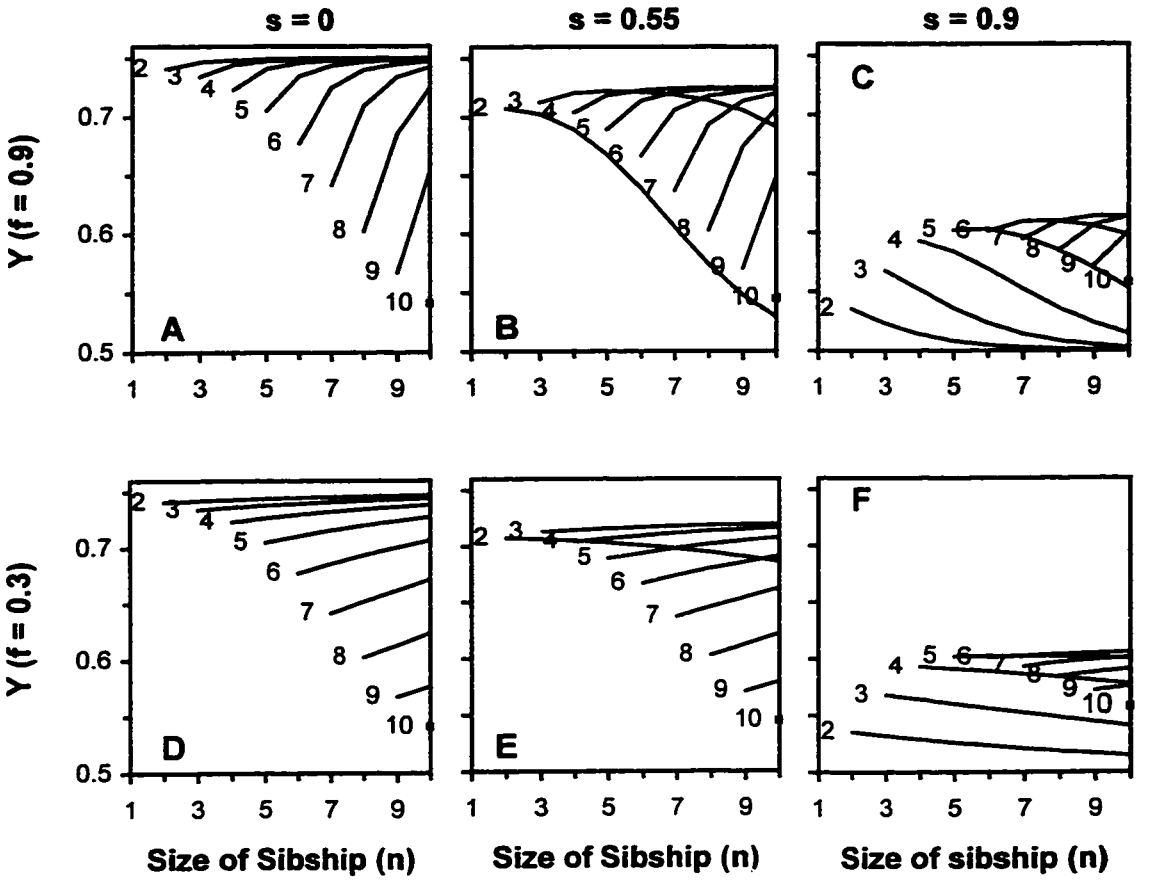


Figure 2.1. Single Locus Models - Effect of penetrance on the expected proportion of alleles shared IBD (Y). All models have a dominant mode of inheritance, with allele frequency (p) equal to 0.01. The numbers in the graph next to the lines indicate the number of affected siblings in the pedigree. f: penetrance of the high-risk allele, s: sporadic proportion.

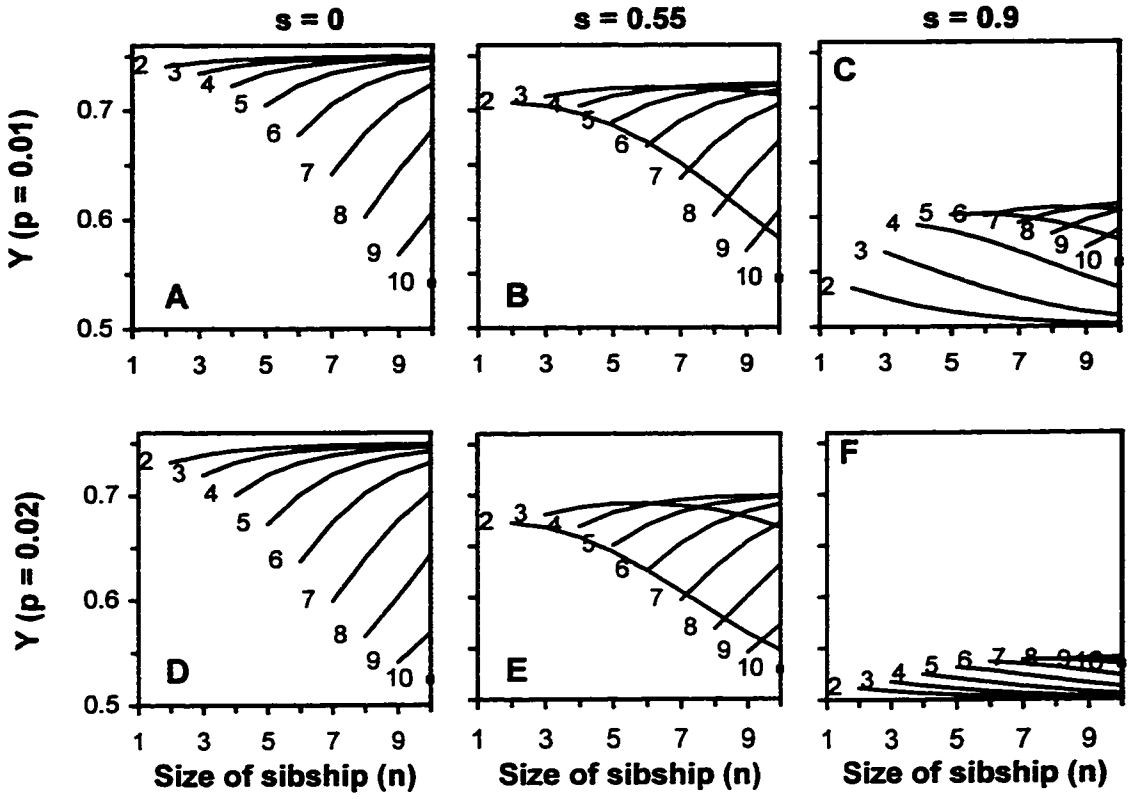


Figure 2.2. Single locus models - effect of allele frequency on Y. All models have a dominant mode of inheritance, with f = 0.7. Other symbols are defined in figure 2.1.

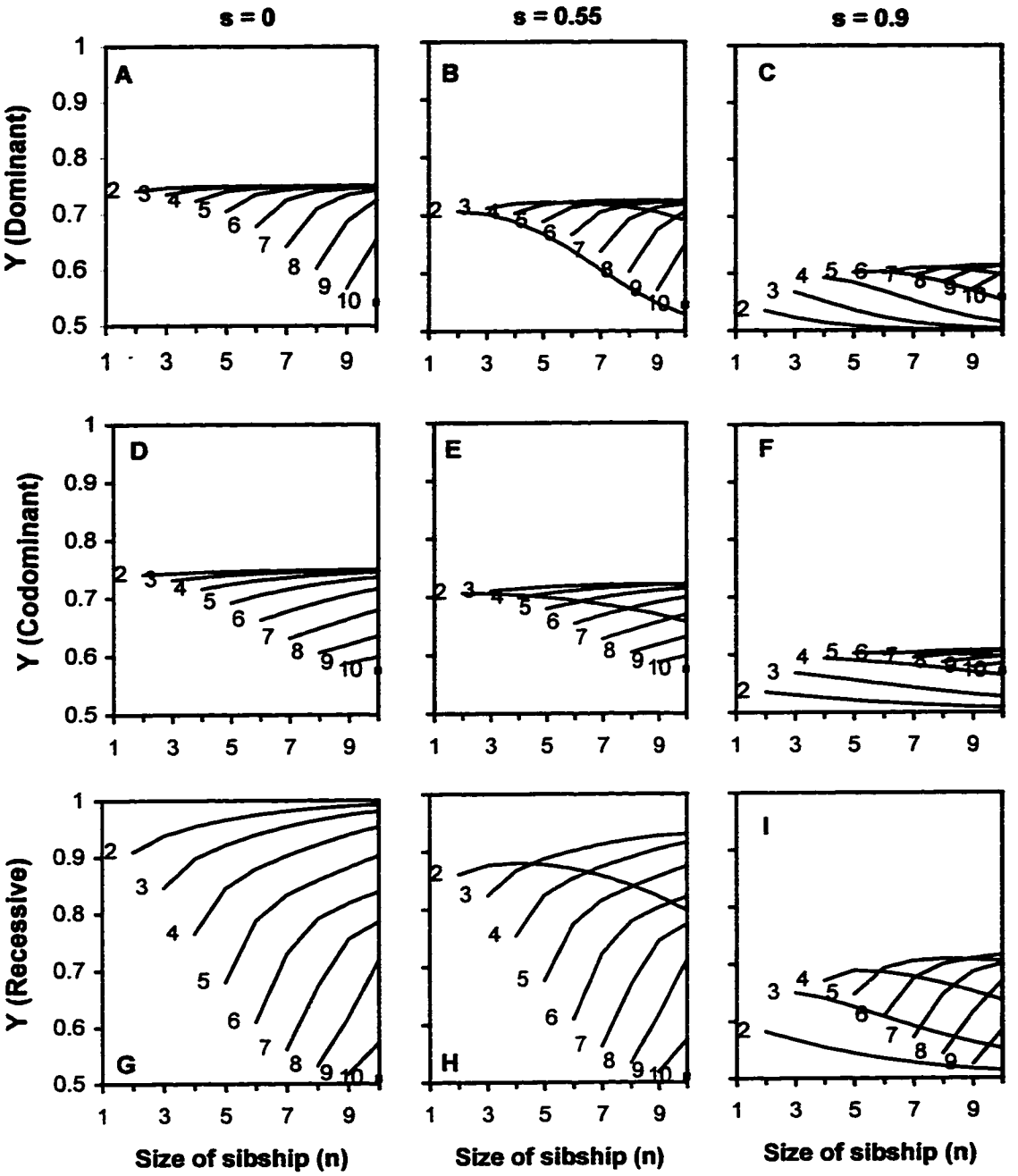


Figure 2.3. Single locus models - effect of mode of inheritance. Dominant model: p = 0.01, f = 0.9, Codominant model: p = 0.01, f_1 = 0.9, f_2 = 0.45, Recessive model: p = 0.1, f = 0.9. s: sporadic proportion. Other symbols are defined in figure 2.1.

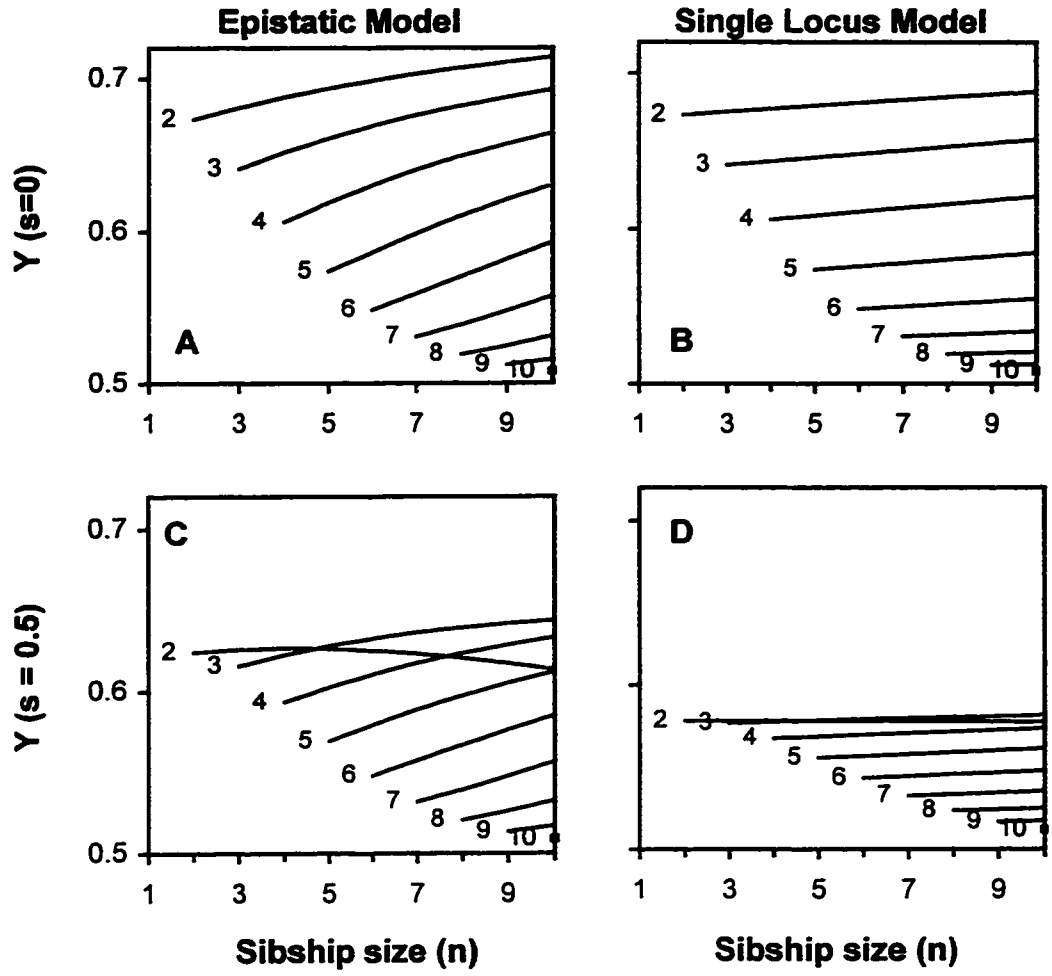


Figure 2.4. Two-locus Epistatic Models. The loci in the epistatic model have the same disease model with a dominant mode of inheritance, p = 0.1, and f = 0.7. Therefore, the value of Y shown here could refer to either locus. The single locus model corresponds to the marginal model for each locus, with a dominant mode of inheritance, p = 0.1, and a penetrance that depends on the sporadic proportion. Other symbols are defined in figure 2.1.

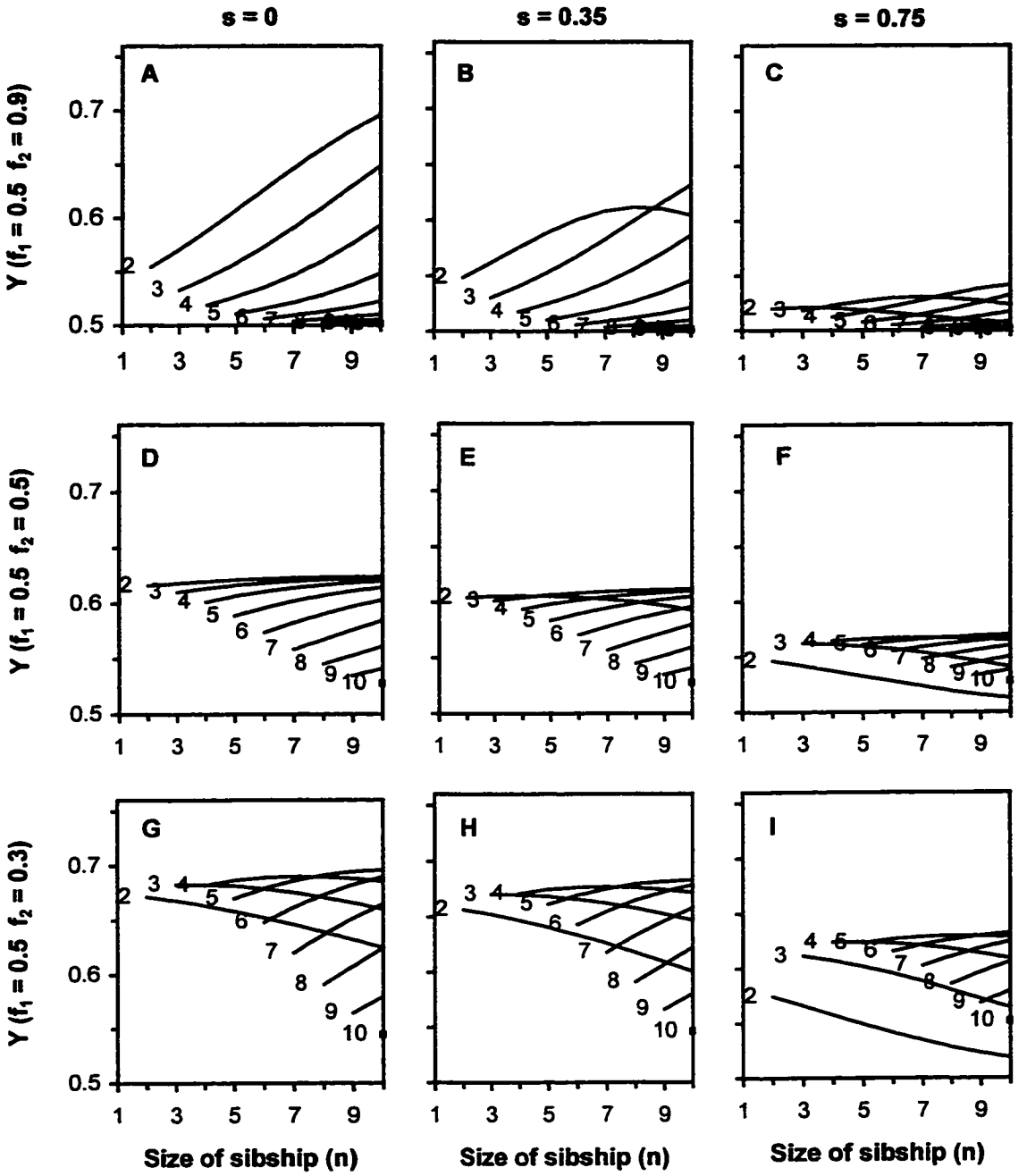


Figure 2.5. Two-locus heterogeneity models - locus 1. Both loci have a dominant mode of inheritance, and p = 0.01. f_1 : penetrance for locus 1, f_2: penetrance for locus 2, s: sporadic proportion. Other symbols are defined in figure 2.1.

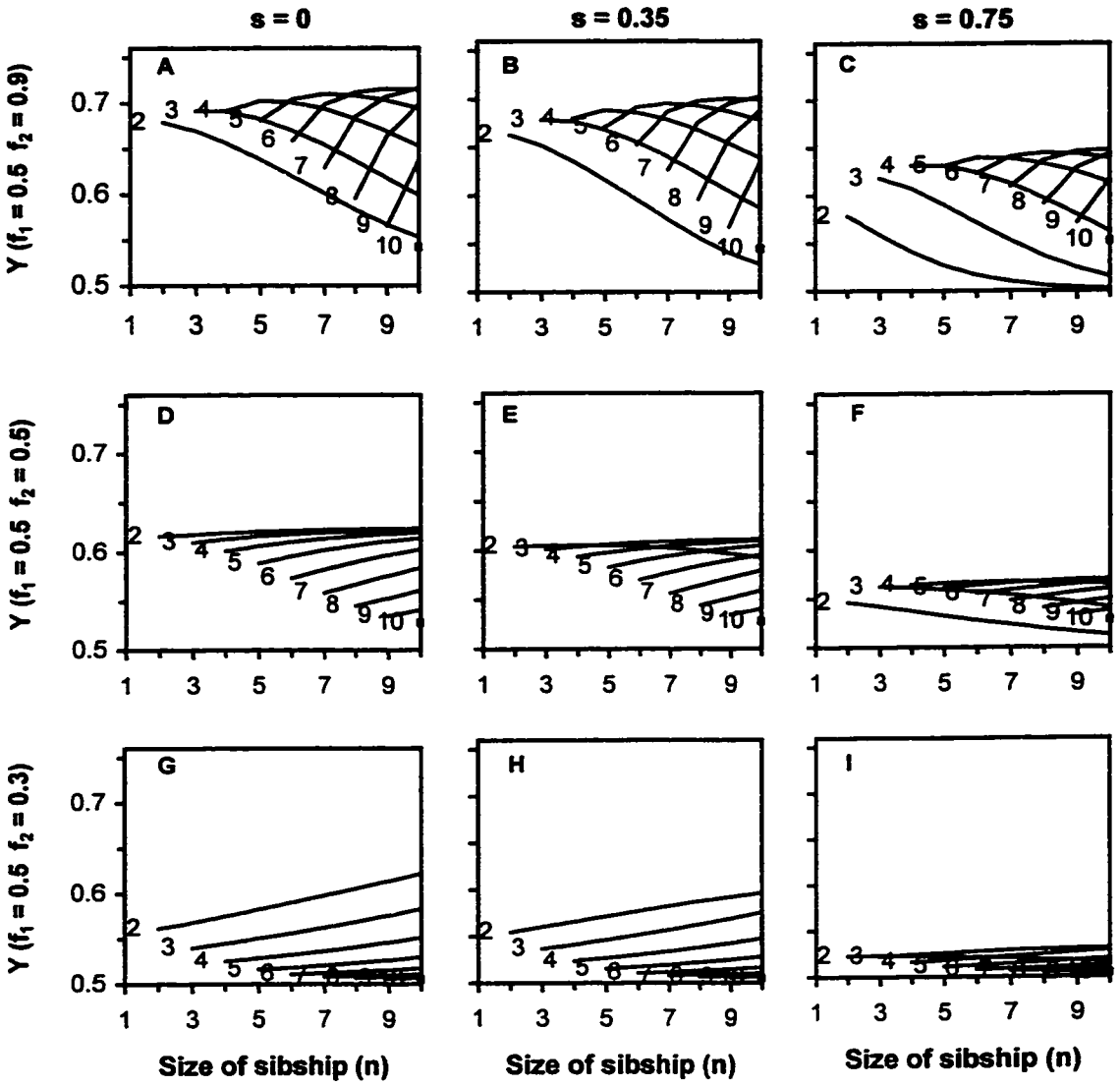


Figure 2.6. Two-locus heterogeneity models - locus 2. Both loci have a dominant mode of inheritance, and p = 0.01. f_1 : penetrance for locus 1, f_2: penetrance for locus 2, s: sporadic proportion. Other symbols are defined in figure 2.1.

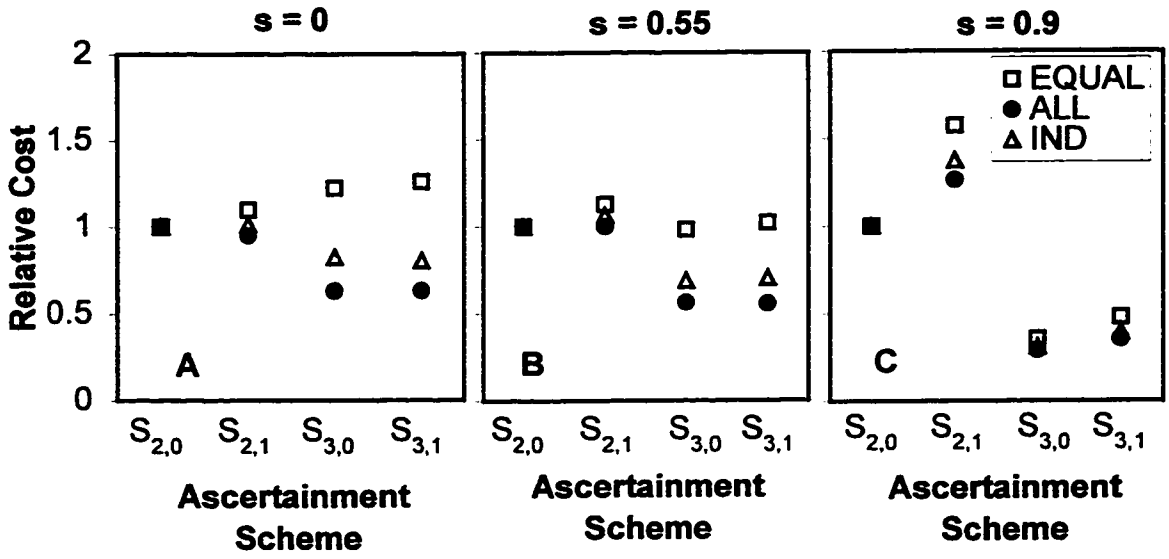


Figure 2.7. Relative cost for single locus models. The models presented here have a dominant mode of inheritance, p = 0.01, and f = 0.9. ALL: all possible ASPs per sibship, IND: all independent ASPs per sibship, EQUAL: one ASP per sibship.

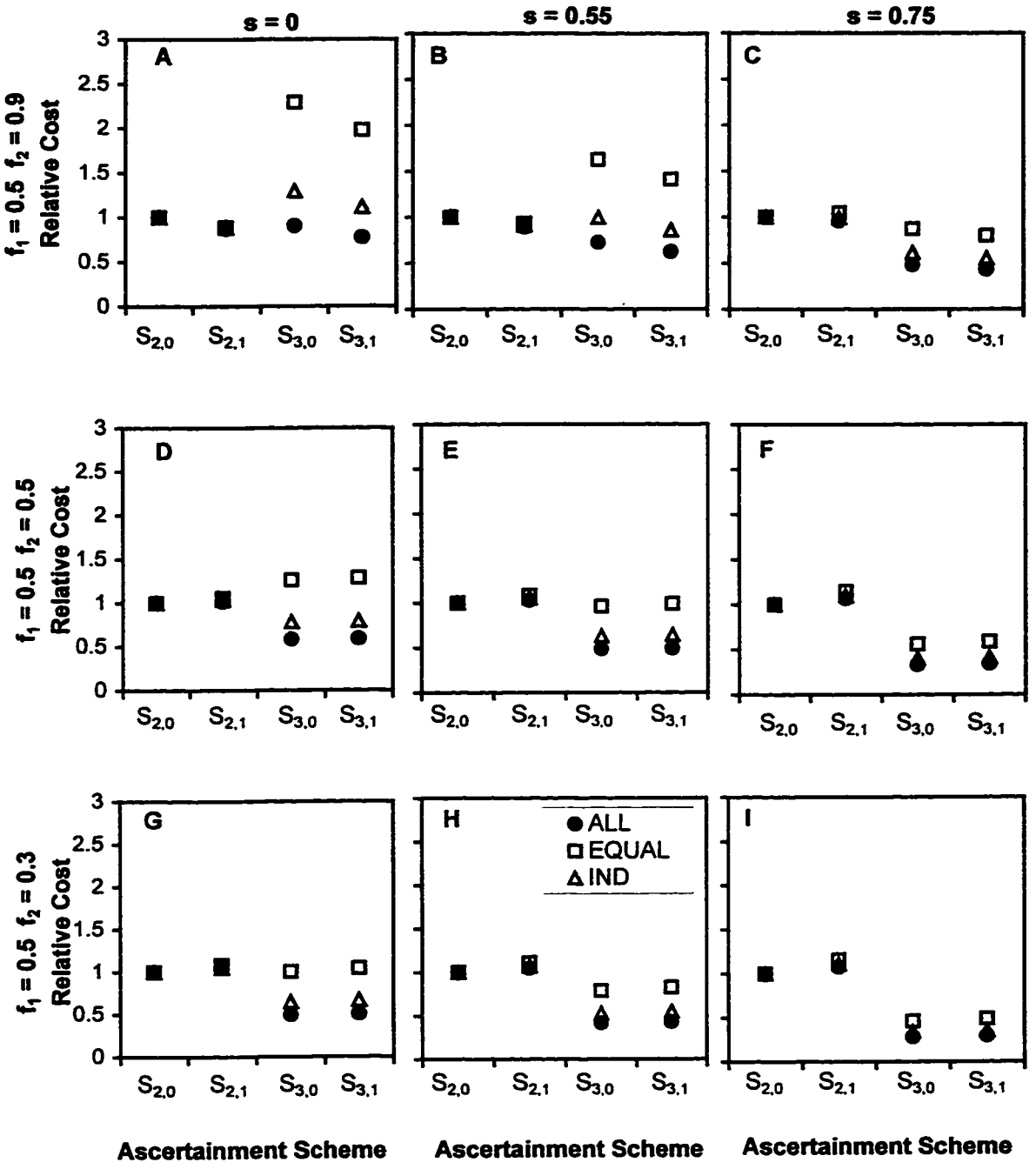


Figure 2.8. Relative cost for two-locus heterogeneity models - locus 1. The models presented here have a dominant mode of inheritance, and p = 0.01 for both loci. f1: penetrance for locus 1, f2: penetrance for locus 2. ALL: all possible ASPs per sibship, IND: all independent ASPs per sibship, EQUAL: one ASP per sibship. Other symbols defined in figure 2.1.

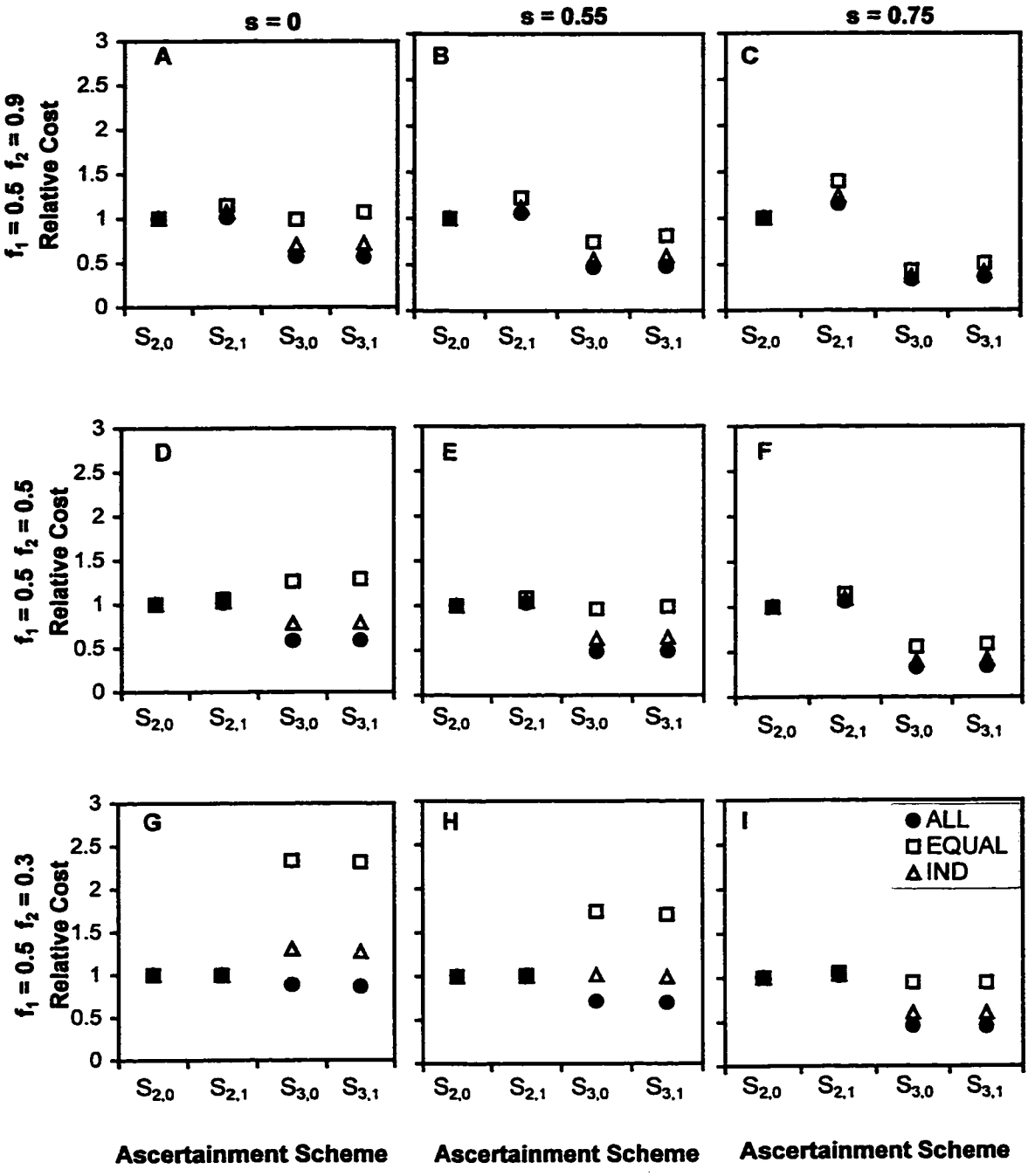


Figure 2.9. Relative cost for two-locus heterogeneity models - locus 2. The models presented here have a dominant mode of inheritance, and p = 0.01 for both loci. f1: penetrance for locus 1, f2: penetrance for locus 2, ALL: all possible ASPs per sibship, IND: all independent ASPs per sibship, EQUAL: one ASP per sibship. Other symbols defined in figure 2.1.

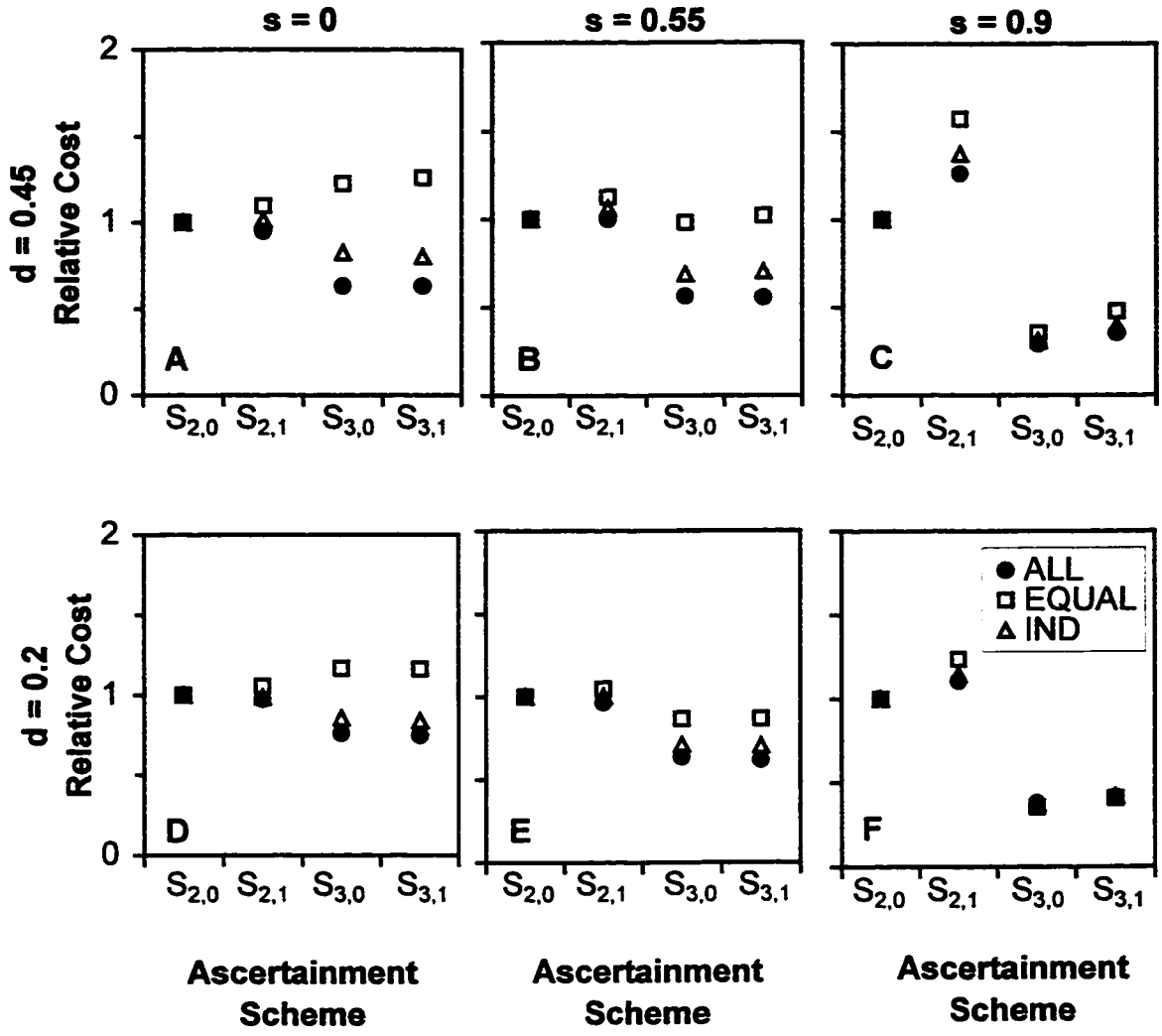


Figure 2.10. Single-locus models - effect of family size distribution. The models presented here have a dominant mode of inheritance, p = 0.01, and f = 0.9. Models in panels A-C used a truncated geometric distribution with parameter d = 0.45 for the distribution of family size, and panels D-F had d = 0.2. ALL: all possible ASPs per sibship, IND: all independent ASPs per sibship, EQUAL: one ASP per sibship. Other symbols defined in figure 2.1.

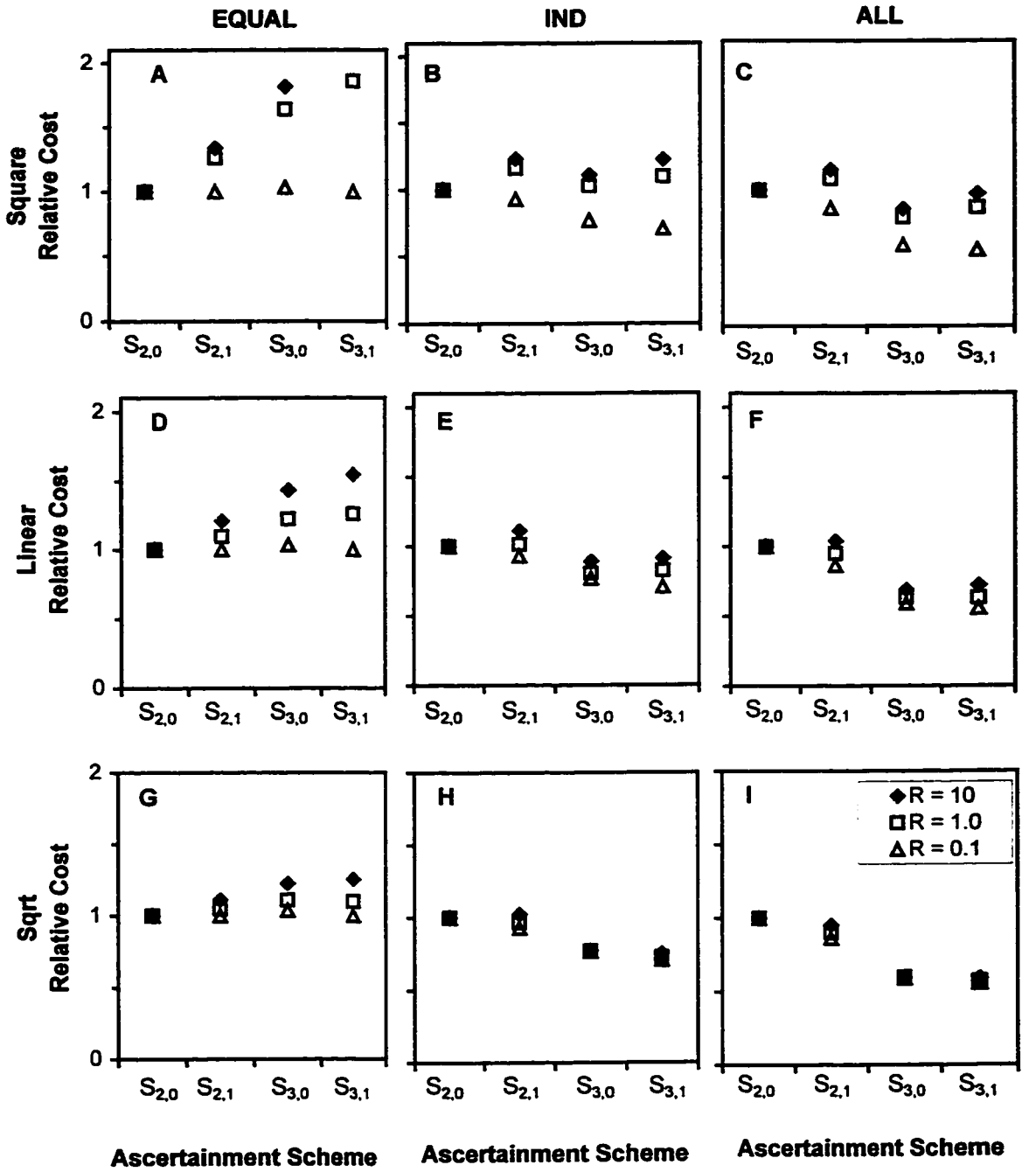


Figure 2.11. Single locus models - effect of cost function. The model presented here has a dominant mode of inheritance, p = 0.01, f = 0.9, and s = 0. ALL: all possible ASPs per sibship, IND: all independent ASPs per sibship, EQUAL: one ASP per sibship. Other symbols defined in figure 2.1.

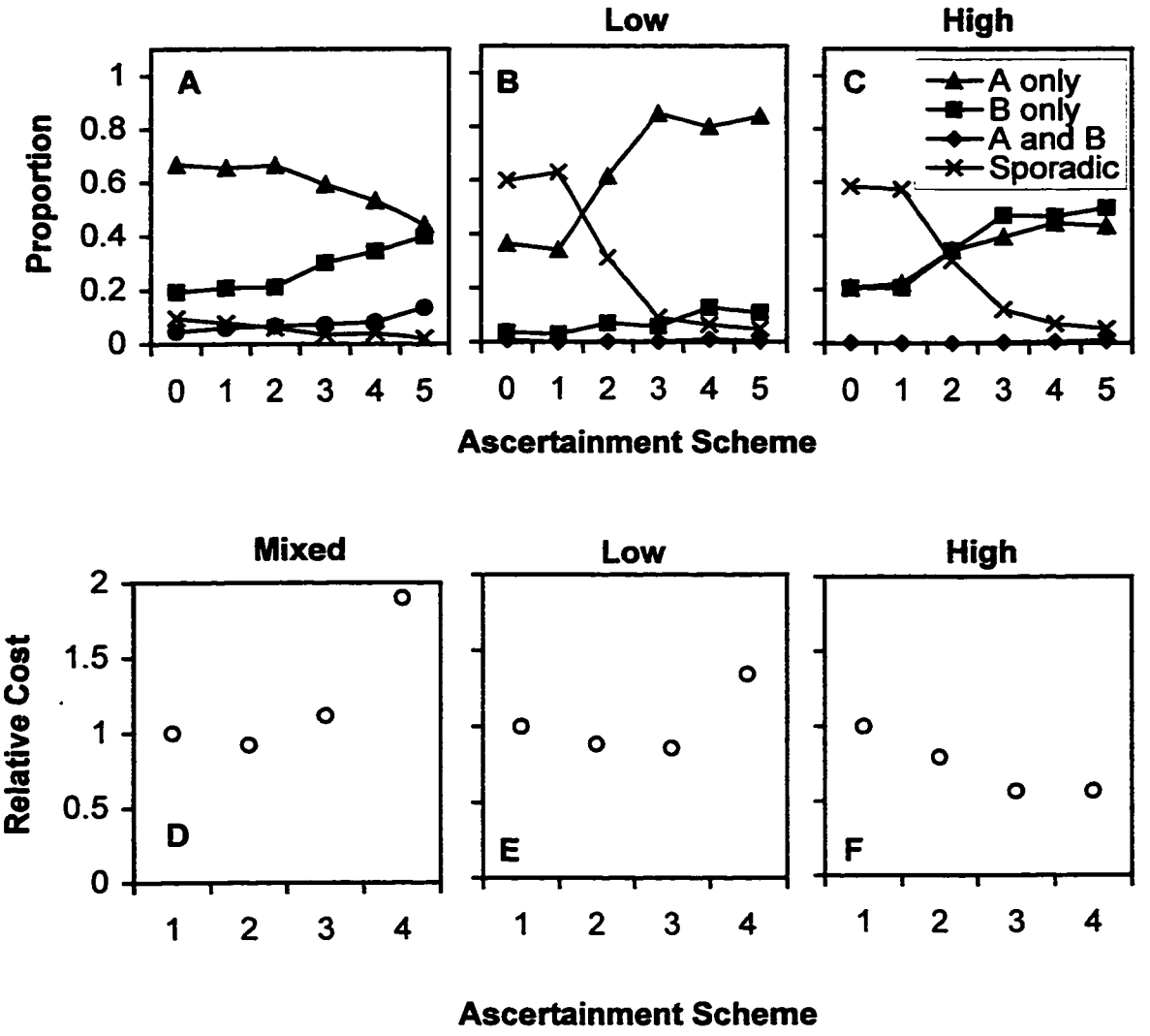


Figure 2.12. Extended Pedigrees. Panels A-C show the proportion of individuals with each form of the disease, A only, B only, A and B, or neither A nor B (sporadic). Panels D-F show the cost of each ascertainment scheme relative to scheme 1. The ascertainment schemes are defined by the minimum number of affected individuals required per pedigree in order to be ascertained. The model parameters for the mixed (M2), low (L3), and high (H4) models are given in table 2.1.

CHAPTER 3 : CHARACTERISTICS OF A GENETIC MAP FOR A COST-EFFECTIVE GENOME SCREEN USING DIALLELIC MARKERS

3.1 INTRODUCTION

Genetic risk factors play an important role in the etiology of many diseases. The identification of genes that contribute to the development of disease helps us to understand the underlying biological mechanisms, and genetic mapping and cloning studies are becoming increasingly popular as a method for identifying such genes. Genetic markers are molecular tools that aid us in this process. Short tandem repeat polymorphisms (STRPs) (Weber and May, 1989) are currently the most commonly used type of marker for genetic mapping studies. A complete genome screen using STRPs usually contains between 200 and 400 markers. There is flexibility in the type of analysis that is performed, and in the pedigree structures that are used for such a screen because the highly informative nature of STRP markers allows extraction of a significant fraction of the available information. Unfortunately, it is difficult to fully automate the genotyping procedure, so it can be relatively expensive to obtain the genotypes with this method.

Many common diseases that potentially have a large public health impact are complex diseases that are known or thought to have a genetic component to the risk. Unfortunately, studies of complex traits often require large sample sizes for genome screens. It is often difficult to accurately estimate the model parameters for complex traits, which results in a reduction in power to detect linkage, and an increased sample size (Clerget-Darpoux et al., 1986; Vieland et al., 1992a,b; Greenberg and Hodge, 1989). Analysis methods that do not require specification of the model are less efficient than correctly specified parametric analyses, so large sample sizes cannot be avoided with these alternative analysis methods (Goldin and Weeks, 1993). Also, complex traits often have multiple genetic and non-genetic forms of the disease, and such heterogeneity within a sam-

ple has also been demonstrated to reduce the power to detect linkage requiring further increases in sample size (e.g. Risch, 1990; Goldin and Weeks, 1993; Goldin and Gershon, 1988). Because of the large sample sizes that are required for complex traits, efforts to reduce the cost of genotyping may significantly impact the cost of such studies.

New technologies have recently been developed for genotyping diallelic markers that may be lower in cost, and more easily automated than current methods. Single nucleotide polymorphisms (SNPs) are diallelic markers that are based on variation in the nucleotide present at single base pairs in the genome. Single base pair substitutions are found approximately every 1-2 kb (Wang et al., 1998), so a large number of markers are potentially available throughout the genome. Methods under development for genotyping SNP markers can potentially be fully automated, which may reduce the cost and the time necessary for obtaining genotypes, while also reducing the error rates (Pease et al., 1994; Nickerson et al., 1990). Unfortunately, SNP markers are relatively uninformative compared to STRPs. This increases the number of markers that must be genotyped to obtain an equivalent amount of information as can be obtained with individual STRPs (Terwilliger et al., 1992). A multipoint method of analysis must be used with SNP markers to obtain this equivalent information. This currently limits the pedigree structures that can be analyzed because of computational constraints, particularly for large, complex pedigrees.

One method to account for the low information available from SNP markers is to identify clusters of closely linked markers. Such a cluster can be treated as a single marker if there is no recombination between the markers within the cluster. The information content of the cluster is similar to that of STRP markers. A second advantage of clustering markers is a more accurate estimate of the recombination fraction between clusters than can be obtained for a uniformly spaced SNP map since for a given sample size, we can estimate large recombination fractions more accurately than small recombination fractions. The effect of map inaccuracy on the power to detect linkage is currently unknown.

Here we address three issues related to the design of cost-effective maps using SNP markers. First, we identify characteristics of the SNP markers that increase the information within the context of maps of clustered SNP markers. Marker characteristics explored here include the allele frequency, the number of loci per cluster, the linkage disequilibrium between markers, and the probability of knowing phase information between pairs of markers. We derive a measure of information for a cluster of markers called the multi-locus PIC (MPIC) to evaluate the effect of these parameters on the information for linkage analysis. We also evaluate uniform vs. clustered marker spacing for SNP markers compared to a map of uniformly spaced STRP markers. We determine the number of markers needed for an equivalent amount of information for each map structure. This allows us to determine the relative cost per marker that is necessary for SNPs to be cost-effective compared to STRPs. Finally, we evaluate the effect of marker distance misspecification on the power to detect linkage, since genetic maps with differences in the marker types and the marker spacing may also differ in the accuracy of the map.

3.2 METHODS

3.2.1 MULTILOCUS POLYMORPHIC INFORMATION CONTENT

The multilocus polymorphic information content (MPIC) is a measure of the information content for a cluster of n diallelic markers, and can be used to evaluate marker characteristics for genetic maps containing clusters of SNP markers. We assume that no recombination occurs between the markers within a cluster. To calculate MPIC, we consider a pedigree with two parents and one offspring. All possible genotype configurations for the trio are listed in table 3.1 for a single marker. Analogous to the calculation for the single-locus PIC (Botstein et al., 1980), here we calculate the probability that we can determine which *haplotype* is transmitted from one parent to the offspring. For the purposes of this discussion, we focus on the transmission from parent 1 to the offspring; however, as with the single-locus PIC, the marker genotypes for parent 2 are also informative in determining the information content for a cluster of markers.

We define a marker within the cluster to be *individually informative* if we can determine which allele at this locus is transmitted from parent 1 to the offspring, and *individually uninformative* if we cannot determine which allele is transmitted. If P(U) is the probability that all loci in the cluster are individually uninformative, and P(I|U) is the probability that the cluster is informative given that all loci in the cluster are individually uninformative, then we can write a general formula for MPIC as:

$$MPIC = 1 - P(U) + P(I|U)P(U).$$

This assumes that the cluster is informative if at least one locus is individually informative, and that it is possible to identify the cases where all of the loci in the cluster are individually uninformative, but the cluster is still informative. We describe these conditions more fully in this section. The probability of each of these cases depends on the allele frequency, the number of loci per cluster, the disequilibrium among loci in the cluster, and the probability of phase information between pairs of loci. These model parameters are described in the next section. A complete formula for MPIC is provided in the appendix.

1. The cluster is informative if at least one locus in the cluster is individually informative. If one locus in the cluster is individually informative, the transmission pattern of the entire haplotype is known since no recombination occurs between the loci in a cluster. Additional informative loci in the cluster only provide information about the phase of the markers within the cluster without providing any additional information about recombination with the trait locus to be mapped. Therefore, the lod score is not affected by additional informative loci in the cluster. The probability that a marker locus is individually informative is equal to the single-locus PIC. Classes C_1 to C_3 from table 3.1 correspond to the cases where the marker is individually informative.

2. If all of the loci in the cluster are individually uninformative, the cluster may still be informative if we have phase information for some pairs of loci.

A pair of marker loci is informative, when each locus is individually uninformative, if we have phase information for all individuals in the parent/offspring trio, and parent 1 is homozygous for one locus, and heterozygous for the other locus. Consider a cluster with only two loci. If parent 1 is homozygous for both loci (class C_5, C_6, or C_7), then the two haplotypes for this individual are identical, and we cannot determine which haplotype is transmitted. If parent 1 is heterozygous for both loci (class C_4), all three individuals in the trio are heterozygous at both loci because the loci are individually uninformative, and the individuals must have the same phase between the two loci because no recombination can occur. In effect, all three individuals are heterozygous for the same two haplotypes, and we cannot determine which haplotype is transmitted. Therefore, in order for a pair of loci to be potentially informative when both loci are individually uninformative, parent 1 must be heterozygous for one locus (class C_4), and homozygous for the other locus (class C_5, C_6, or C_7). One additional requirement for the pair of loci to be informative is that we have phase information for all three individuals in the trio. Without phase information for all three individuals, either allele can be transmitted from parent 1 to the child. The probability that we know phase information is discussed in section 3.2.1.1 below. Note that loci in class C_8 (all three individuals are homozygous for the same two alleles) are never informative, even in combination with other loci, because all of the alleles are the same for all individuals.

3. If all of the loci are individually uninformative, we do not need to consider more than pairs of loci.

Consider a cluster of three loci where all the loci are individually uninformative, and none of the pairs of loci are informative. Parent 1 cannot be either homozygous or heterozygous for all three loci for the reasons stated above in part 2. Therefore, we only need to consider the case where parent 1 has both homozygous and heterozygous loci (class C_4 – C_7). Since all of the loci in the cluster are individually uninformative, we need phase information for the cluster to be informative. If we do have phase information, then it is sufficient to only have information for a pair of loci as described in part 2, so we do not need to consider phase information for all three loci. A similar argument holds for clus-

ters with more than three loci. Therefore, it is not necessary to consider more than pairs of loci to determine cases where the cluster is potentially informative when all of the loci are individually uninformative.

3.2.1.1 Model parameters

Now we consider model parameters that determine the value of MPIC including the allele frequency, the number of loci per cluster, the disequilibrium between loci, and the probability of phase information for pairs of loci. First, we consider the effect of the marker allele frequency on the value of MPIC. At each marker locus in the cluster there are two alleles with frequencies p_i and $q_i = 1 - p_i$ for marker *i*. We assume that p_i refers to the more common of the two alleles, so it ranges between 0.5 and 1.0. We also evaluate the effect of the number of loci per cluster (n) on the value of MPIC, where *n* ranges between 2 and 10 loci per cluster.

Another parameter that may affect the value of MPIC is the probability that we know the phase between loci *j* and *k* for individual *i* (π_{ijk}). We assume that $\pi = \pi_{ijk}$ is the same for all individuals, and for all pairs of loci in a cluster. In addition, we assume that the pairs of loci in the cluster are independent in terms of which pairs have phase information. These simplifying assumptions are generally incorrect, but do not have a large impact on the value of MPIC as we will show. Phase information can be obtained by several methods. First, phase information for a pair of loci can be identified from the genotypes of other pedigree members. This suggests that larger pedigrees tend to have higher values of π than smaller pedigrees. Secondly, molecular methods can be used to obtain phase information, although this method would increase the cost of the analysis.

The final parameter of interest is the linkage disequilibrium (δ) present between loci in the cluster. The euclidean distance is a composite measure of disequilibrium for the entire cluster:

$$\delta = \sqrt{\sum_{i=1}^{2^n} (P_i - Q_i)^2},$$

where P_i is the observed haplotype frequency for haplotype *i*, and Q_i is the expected haplotype frequency under the assumption of linkage equilibrium. When the euclidean distance is equal to 0 there is no disequilibrium, and for values significantly greater than 0 there is disequilibrium. Note that the δ parameter is not actually used in the calculation of MPIC, it is simply a measure of the composite disequilibrium for the cluster. However, this distance measure is useful for evaluating the effect of linkage disequilibrium on the information for a cluster because it summarizes the difference between the observed haplotype frequencies and the frequencies expected under linkage equilibrium. There are an infinite number of possible ways that the observed haplotype frequencies can differ from the frequencies expected under linkage equilibrium. MPIC is calculated for a sample of the possibilities that cover the possible range of the euclidean distance.

3.2.2 Comparison of genetic map designs

The uniformly spaced and clustered SNP map structures and the uniformly spaced STRP map structure (figure 3.1) were compared with simulation methods. Figure 3.2 depicts the four pedigrees used to evaluate the map structures. The disease model was a single-locus model with a dominant mode of inheritance, complete penetrance, and an allele frequency of 0.001 for the high-risk allele. The STRP map structure consisted of markers with four equally frequent alleles. Such a marker is representative of a typical STRP marker since the PIC is 0.7, while the median PIC from a representative data set is 0.68 (version 9 screening set, http://www.marshmed.org/ genetics; Broman et al., 1998). The SNP genetic map structures consisted of markers with equal allele frequencies. A 100 cM chromosome was simulated for each map structure with marker spacing between 1 and 12 cM. The disease locus was located halfway between two markers approximately one third of the total length from one end of the chromosome. One hundred data sets with 10 pedigrees each were simulated for each marker type and marker spacing combination.

GENEHUNTER (Kruglyak et al., 1996) was used to calculate the information, which was measured halfway between the two markers in the middle of the chromosome. The measure of information calculated in GENEHUNTER is based on the entropy of the possible inheritance vectors (i.e. the transmission pattern) for the pedigree. If the inheritance vector is known exactly, the information is the maximum value of 1.0, and as the number of possible inheritance vectors increases, the information decreases.

For each map structure, the relationship between the information (I) and the marker spacing (S) was determined by linear regression from the observed values given by the simulation described above. These results are used to determine the relative cost of map structures compared to a uniform STRP map structure in terms of the relative number of markers that are required for each structure. For map structure *j*, the number of markers necessary for a given level of information (n_{Ij}) is determined from the total chromosome length (100 cM), the marker spacing required for information I (S_I), and the number of markers per cluster (n = 1 for uniformly spaced markers) as:

$$n_{Ij} = 100n/S_I.$$

The information was evaluated for values that correspond to a uniform STRP map structure with marker spacing between 5 and 10 cM. The cost per person of performing a genome screen depends on the number of markers required, and the cost of genotyping per marker. Therefore, the relative cost of map structures *j* and *k* ($C_{j,k}$) is proportional to the ratio of the number of markers required for each map structure, $C_{j,k} \propto n_{Ij} / n_{Ik}$.

3.2.3 GENETIC MAP ACCURACY

To evaluate the effect of marker distance misspecification, we considered a single interval between two markers. If the disease locus was linked to the marker loci, it was located halfway between the markers as depicted in Figure 3.3. θ_1 is the recombination fraction between marker 1 (M_1) and the disease locus (D), θ_2 is the recombination fraction between marker 2 (M_2) and the disease (D), and θ_{12} is the recombination fraction

between the two marker loci, M_1 and M_2. $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_{12}$ are the estimated values of θ_1, θ_2, and θ_{12} respectively. If the disease locus was unlinked, we still considered the estimate of the recombination fraction between the two marker loci, $\hat{\theta}_{12}$. The effect of marker distance misspecification on the lod score was evaluated using the bias in the expected maximum lod score (Emlod). The bias is equal to:

$$\text{bias} = \text{Emlod}(\hat{\theta}_{12}) - \text{Emlod}(\theta_{12}),$$

where the Emlod(θ) is the expected maximum lod score at a recombination fraction of θ.

We estimated the Emlod using simulation. Twenty nuclear pedigrees with three offspring each were simulated per data set, giving a total of 40 potentially informative meioses per data set. Two types of markers were simulated: a marker with 20 equally frequent alleles (high information content), and a diallelic marker with two equally frequent alleles (low information content). The disease locus model had a single locus with a dominant mode of inheritance, complete penetrance, and an allele frequency of 0.01 for the high-risk allele. 500 data sets were simulated for each value of θ_{12} including θ_{12} = 0.02, 0.04, 0.05, and 0.1. Each data set was analyzed using GENEHUNTER (Kruglyak et al., 1996) at 35 values of $\hat{\theta}_{12}$ including increments of 0.002 for $\hat{\theta}_{12}$ between 0 and 0.05, and increments of 0.005 for $\hat{\theta}_{12}$ between 0.05 and 0.1. The lod score was calculated at four points in the interval between the two markers for each value of $\hat{\theta}_{12}$ and at each marker locus. The maximum lod score for each data set was the maximum lod score out of these six points. The Emlod was estimated by averaging the maximum lod scores for all data sets.

We determine the expected bias when θ_{12} is underestimated for low and high information content markers. If x is the number of recombinants that are observed out of n informative meioses, then x is distributed as a binomial(n,θ_{12}). θ_{12} is underestimated when $x/n = \hat{\theta}_{12}$ is less than θ_{12}. Therefore, we can find the expected bias by summing over all values

of x that satisfy $x/n = \hat{\theta}_{12} < \theta_{12}$, such that the expected bias equals $\frac{\sum_x P(x)bias(x/n)}{\sum_x P(x)}$, where $P(x)$ is the binomial probability of observing x.

3.3 RESULTS

3.3.1 Effect of model parameters on MPIC

There is an inverse relationship between allele frequency and MPIC, although the decrease in MPIC is generally small for allele frequencies between 0.5 and 0.75 (Figure 3.4). For example, a change in allele frequency from 0.5 to 0.75 causes a decrease in MPIC from 0.98 to 0.94 with 6 markers per cluster, and from 0.70 to 0.56 (the worst case) with 2 markers per cluster. This indicates that marker allele frequencies can differ substantially from the ideal case of equal frequencies and still be informative in the context of clustered sets of markers. This is important since the allele frequencies may not be the same in all populations, so it would be impossible to obtain a single marker set that is ideal for every population. For values of π less than 1.0, the relationship between allele frequency and MPIC is similar (data not shown).

As expected, the value of MPIC increases with additional loci in the cluster, where the largest increase occurs between one and two loci per cluster (Figure 3.5). For example, when the allele frequency is 0.5 for all loci in the cluster, an increase from one to two loci corresponds to an increase in MPIC from 0.375 to 0.70, while an increase from five to six loci corresponds to an increase in MPIC from 0.97 to 0.98. For allele frequencies in the range of interest (0.5 – 0.75), MPIC does not substantially increase with more than five markers per cluster, suggesting that we only need to consider genetic map structures with at most five loci per cluster. For values of π less than 1.0, the relationship between the number of loci and the information is similar (data not shown).

The effect of phase information on the value of MPIC is relatively small compared to the other parameters considered here. There is only a slight increase in the information as the

phase information increases (Figure 3.6). For example, for a change in π from 0 to 1, MPIC increases from 0.61 to 0.70 for two loci per cluster, and from 0.94 to 0.98 for six loci per cluster. These differences are much smaller than the differences observed across the entire range of the other parameters. In calculating MPIC, we assume that π is the same for all individuals, and for all pairs of loci in the cluster. Correcting this assumption might change the shape of the curve, but would not have an effect on the absolute difference in the value of MPIC between $\pi = 0$ and $\pi = 1$. These are the maximum and minimum possible values of MPIC, when phase is known for all individuals, and for all pairs of loci ($\pi = 1$), and when phase is not known for all individuals, and for all pairs of loci ($\pi = 0$). Therefore, this assumption does not have a large effect on the results, since the change in the value of MPIC due to the probability of knowing phase information is small.

Finally, the presence of linkage disequilibrium among loci in a cluster generally reduces the value of MPIC, although modest amounts of linkage disequilibrium do not significantly reduce the information (Figure 3.7). Linkage disequilibrium can increase the value of MPIC in some cases if the haplotype frequencies are more uniform under disequilibrium than the expected haplotype frequencies under equilibrium. However, as in the example presented here, if the allele frequencies are exactly equal for all markers in the cluster, the haplotype frequencies are uniform under equilibrium, and therefore linkage disequilibrium will always reduce the information. When the markers are in complete linkage disequilibrium (only two haplotypes are observed), the value of MPIC corresponds to the value of the single-locus PIC for a marker with two alleles. In the example, all of the loci have equal allele frequencies, so the single-locus PIC is equal to 0.375 (Figure 3.7). In general, the value of MPIC is only slightly decreased when the euclidean distance is between 0 and 0.2. The exact haplotype frequencies under disequilibrium that correspond to this range of the euclidean distance depend on the number of markers in the cluster, and the allele frequencies for the markers. As an example, a cluster of three markers with equal allele frequencies has haplotype frequencies equal to 0.125 under

equilibrium. In this case, haplotype frequencies between 0 and 0.2 under disequilibrium correspond to the appropriate range of the euclidean distance.

3.3.2 Comparison of genetic map structures

Figure 3.8 depicts the relationship between marker spacing and information for four pedigree structures. For a given marker spacing, the uniformly spaced SNP map structure has the lowest information, since the individual SNP markers are less informative than clusters of SNPs or individual STRPs. The uniformly spaced STRP map structure and the map structure with two SNPs per cluster have similar values of information for a given marker spacing. A cluster of two SNPs with complete phase information is equivalent to a four-allele STRP, so the information should be similar for these two map structures. The information for a given marker spacing increases with the number of markers per cluster. Therefore, clusters with fewer markers require a smaller marker spacing to achieve the same information content as clusters with more markers. These results are consistent for all of the pedigree structures evaluated.

In general, the genetic map structures using SNPs require more markers than the STRP map structure, which has an impact on the relative cost per genotype that is necessary for SNPs to be cost-effective (Table 3.2). The most cost-effective SNP map structure is the uniform structure, which requires approximately 1.8 times the number of markers required for the uniform STRP map structure. This means that a 6 cM (3 cM) uniform SNP map structure is approximately equivalent to a 10 cM (5 cM) uniform STRP map structure. The clustered SNP map structures require approximately 2.0, 2.5, and 3.0 times the number of markers required for the STRP map structure for 2, 3, and 4 markers per cluster respectively. All of the clustered map structures require more markers than the uniform SNP map structure, although the map structure with two SNPs per cluster requires only slightly more than the uniform SNP map structure. These results indicate that SNP markers must cost less per marker than STRP markers to be cost-effective since the SNP map structures always require more markers than the STRP map structure.

3.3.3 GENETIC MAP ACCURACY

In the case where the disease locus is linked to the markers, the lod score is reduced if the marker distance is misspecified as might be expected, which decreases the power to detect linkage (Figure 3.9). The bias in the lod score is generally greater if the marker distance is underestimated than if it is overestimated. In addition, we found that the bias is larger for less informative markers than for more informative markers. This result can be explained by considering whether neither, one, or both marker loci are informative for a given meiosis. If neither locus is informative, the meiosis does not contribute to the lod score. If only one locus is informative, the bias in the lod score is greater than if both loci are informative. Less informative markers are more likely to have only one informative locus than more informative markers; therefore, the magnitude of the bias is generally greater for less informative markers. All of the results presented here are dependent on the sample size in that the magnitude of the bias is smaller when the sample size is reduced (i.e. if there are no informative meioses, there is no information, so the magnitude of the bias goes to zero). However, the general conclusions about the direction of the bias should not depend on the sample size.

If the trait locus is unlinked, the lod score is reduced if the marker distance is underestimated, but is inflated if the marker distance is overestimated (Figure 3.10). The magnitude of the bias is smaller for the low information content markers than for the high information content markers, but the direction of the bias is the same in both cases. The magnitude of the bias is much greater in the unlinked case than in the linked case for a given sample size. In the example in figure 3.10 ($\theta_{12} = 0.05$), the magnitude of the bias in the linked case ranges between 0 and –0.6, while in the unlinked case the magnitude of the bias ranges between –24.1 and 5.6.

For a given sample size, the estimate of the recombination fraction is more accurate for high information content markers than for low information content markers, which affects the expected bias in the lod score. As an example we consider the magnitude of the expected bias for low and high information content markers when the recombination

fraction is underestimated in the linked case (Table 3.3). There are generally two samples of pedigrees that are used for genetic mapping, the mapping sample (e.g. the CEPH pedigrees) is used to estimate the recombination fraction between markers, and the disease sample (e.g. the pedigrees collected for the current study) is used to map the trait locus. If there are 200 potentially informative meioses available in the mapping sample, the expected bias for the high information content markers is approximately –0.01. For the low information content markers, not all of the meioses are informative for linkage, so 200 potentially informative meioses produces only 28 informative meioses in the mapping sample. In this case, the expected bias is between –0.16 and –0.26 in this case, which is much greater than the bias for the high information content markers. A sample size of 200 was chosen for this example because this is similar to the sample size that is currently used to estimate marker distances in some groups (Broman et al., 1998). It is interesting to note that a sample size of approximately 1400 potentially informative meioses would be needed to produce 200 informative meioses for the low information content markers, where the expected bias of approximately –0.03 is still greater than the expected bias for the high information content markers. These results indicate that using more informative markers in the genetic map will minimize the expected bias when the marker distance is underestimated.

3.4 DISCUSSION

We have identified several characteristics of SNP markers that are required for a cost-effective genome screen. First, we derived a statistic called MPIC for calculating the information content of a cluster of diallelic markers that is analogous to the single locus PIC. Using MPIC, we evaluated marker characteristics for clustered SNPs to identify a desirable range of each parameter. We found that the value of MPIC is approximately the same for allele frequencies of the common allele in the range of 0.5 and 0.75. Kruglyak (1997) suggested that allele frequencies for uniformly spaced SNP map structures should be between 0.5 and 0.8, which is similar to our results for clustered SNP map structures. In addition, there can be some linkage disequilibrium present between the loci in the

cluster without a significant reduction in the information. We found that the value of MPIC does not significantly increase if there are more than five markers per cluster, suggesting that we should only consider map structures with at most five markers per cluster. Finally, there is only a slight increase in the value of MPIC with an increase in the probability of phase information, so it is not important to accurately estimate this parameter.

In the ideal case where the marker distances are known, we conclude that clustered SNP map structures are less efficient than the uniform SNP map structure based on the number of markers required for each map structure. Not surprisingly, we found that all of the SNP map structures require more markers than the STRP map structure, and that SNPs can cost at most 60% of the cost per genotype for STRPs to be cost-effective. It is unclear whether the technologies under development for genotyping SNP markers can achieve this reduction in cost. The marker spacing for the uniform SNP map structure of 6 cM (3 cM) is slightly larger than the spacing suggested by Kruglyak (1997), where a 4.5 cM (2 cM) uniform SNP design was equivalent to a 10 cM (5 cM) uniform STRP design. If we do consider the effect of misspecifying the marker distance, we found that using more informative markers such as the STRP markers or clusters of SNP markers will minimize the expected bias in the lod score when the markers are linked to the disease locus. Therefore, a better strategy might be to use a genetic map with clusters of two SNP markers since this will reduce the bias in the lod score while only slightly increasing the cost compared to uniformly spaced SNP markers.

One limitation to our approach is that we only considered the effect of misspecifying a single marker interval on the lod score. However, for uniform or clustered SNP map structures, we may need to consider multipoint analyses with more than two markers since the markers are individually relatively uninformative. Although the effect of misspecification of multiple marker intervals is unknown, it may be much greater than the magnitude of the bias for a single interval. Evaluation of this more complex situation is more difficult because of the large number of model parameters that need to be evaluated.

The fact that multipoint analysis methods must be used is an important consideration in evaluating the utility of designs based on use of SNP markers. Although most evaluations of multipoint linkage analysis have been done under the assumption of a perfect map, there is increasing power to detect linkage with an increasing number of markers in the analysis (Amos et al., 1997). Unfortunately, multipoint linkage analysis is not as robust to model misspecification as two-point linkage analysis (Risch and Giuffra, 1992). In addition, the current methods available for computing exact multipoint likelihoods are limited due to computational constraints either in the number of markers that can be considered simultaneously (e.g. VITESSE, O'Connell and Weeks, 1995), or in the pedigree structures that can be evaluated (e.g. GENEHUNTER, Kruglyak et al., 1996). The clustered design offers some flexibility in analysis methods since it may not be necessary to simultaneously consider all of the markers on the chromosome. However, it is still necessary to consider all of the markers within a cluster, which may not be possible for all pedigree structures. In particular, pedigrees with inbreeding loops, or with multiple founder sets may be impossible to evaluate within the available resources. Other methods currently under development include variance component methods (Almasy and Blangero, 1998) and Markov chain Monte Carlo (MCMC) methods (Heath, 1997). The MCMC method can be used with more extended and complex pedigrees with many markers. However, these methods are computationally intensive as well, and current implementations are limited in the kinds of phenotypes that can be evaluated.

Another factor to consider in evaluating map structures using SNP markers is that particular diallelic polymorphisms may differ in the model parameters such as the allele frequency and the linkage disequilibrium that are found in different populations. For map structures using SNPs it may be necessary to incorporate redundancy into the screening set such that there is more than one marker representing each region of the genome. This would increase the chance that at least one marker per region is informative for linkage; however, this also reduces the cost per genotype that is necessary for a map structure based on SNPs to be cost-effective compared to STRP markers. This problem is some-

what reduced by the fact that a range of possible values of the allele frequency and the linkage disequilibrium are essentially equivalent in information content.

When all of these factors are considered, a clustered map structure with two markers per cluster may provide the most flexible alternative for using SNP markers, while only slightly increasing the cost over the most cost-effective strategy of uniform spacing. A map of clustered pairs of SNPs at 10 cM spacing has approximately the same information content as a map of uniformly spaced STRP markers with 10 cM spacing so the bias from marker distance misspecification is approximately the same in these two cases. In addition, the clusters can be used individually instead of in a multipoint analysis, so there is more flexibility in the type of analysis, and in the pedigree structures that can be considered.

3.5 APPENDIX

In this appendix, we show the exact formula for MPIC. MPIC can generally be written as:

$$\text{MPIC} = 1 - \text{P(U)} + \text{P(I|U)P(U)},$$

where P(U) is the probability that all of the loci are individually uninformative, and P(I|U) is the probability that the cluster is informative given that all of the loci in the cluster are individually uninformative. Assume there are n loci in the cluster, and $w = 2^n$ possible haplotypes. The haplotypes, denoted $h_1, \ldots, h_w$, have frequencies $p_1, \ldots, p_w$, where each haplotype is a vector of length n, which corresponds to the multi-locus genotype. π is the probability that we know the phase between two loci, and we assume this probability is the same for all individuals, and all pairs of loci. To calculate the first term, P(U), we sum over all possible pairs of haplotypes for each parent (haplotypes i and j for parent 1, and haplotypes k and l for parent 2) and the offspring haplotypes given the parental haplotypes (haplotype i or j from parent 1, and haplotype k or l from parent 2). The

frequency of each term is only included if all loci are individually uninformative. This term is equal to:

$$P(U) = \sum_{i=1}^{w}\sum_{j=1}^{w}\sum_{k=1}^{w}\sum_{l=1}^{w}\sum_{r=\{i,j\}}\sum_{s=\{k,l\}}\frac{p_i p_j p_k p_l}{4}\prod_{t=1}^{n} 1_{\left(h_{it}h_{jt},h_{kt}h_{lt},h_{rt}h_{st}\right)},$$

Where $p_i p_j p_k p_l$ is the probability of the parental haplotypes, ¼ is the probability of each pair of haplotypes for the offspring, and

$$\left(h_{it}h_{jt},h_{kt}h_{lt},h_{rt}h_{st}\right) = \begin{cases} 0 & \text{if locus } t \text{ is informative} \\ 1 & \text{if locus } t \text{ is uninformative} \end{cases}.$$

Locus t is informative if $h_{it}h_{jt}$ (the genotype for parent 1 at locus t) is heterozygous, and if either $h_{kt}h_{lt}$ (the genotype for parent 2 at locus t) or $h_{rt}h_{st}$ (the genotype for the offspring at locus t) is homozygous. The product over all loci is equal to 1 if all of the loci are individually uninformative, and equal to 0 if at least one locus in the cluster is individually informative.

For the second term, P(I|U), the cluster is informative if we know the phase for at least one pair of potentially informative loci in the cluster. Therefore, the probability that the cluster is informative is one minus the probability that we do not know the phase for any of the potentially informative pairs. There are three types of potentially informative pairs based on the class of the loci in the cluster (all classes are listed in table 2.1). The number of loci in each class of interest, denoted c_4, c_5, c_6, or c_7, can simply be counted given the haplotypes of the parents and the offspring. For pairs of loci where one locus is in class C_4, and the other locus is in class C_6, the pair is informative if we know the phase for parent 2. The probability that all pairs of loci of this type are uninformative is $(1-\pi)^{c_4 c_6}$. For pairs of loci where one locus is in class C_4, and the other locus is in class C_7, the pair is informative if we know the phase for the offspring. The probability that all pairs of loci of this type are uninformative is $(1-\pi)^{c_4 c_7}$. Finally, pairs of loci where one locus is in

class C_4, and the other locus is in class C_5, the pair is informative if we know the phase for either the offspring or parent 2. Therefore, the probability that all pairs of loci of this type are uninformative is $\left(1-2\pi+\pi^2\right)^{c_4c_5}$. The probability that the cluster is informative given that all of the loci are individually uninformative for particular haplotypes of the parent-offspring trio is

$$b_{ijklrs} = 1-(1-\pi)^{c_4c_6}(1-\pi)^{c_4c_7}\left(1-2\pi+\pi^2\right)^{c_4c_5}.$$

When we combine these two terms, the value of MPIC is

$$\text{MPIC} = 1-\sum_{i=1}^{w}\sum_{j=1}^{w}\sum_{k=1}^{w}\sum_{l=1}^{w}\sum_{r=\{i,j\}}\sum_{s=\{k,l\}}\frac{p_i p_j p_k p_l}{4}\prod_{t=1}^{n}1_{\left(h_{it}h_{jt},h_{kt}h_{lt},h_{rt}h_{st}\right)}\{1-b_{ijklrs}\}.$$

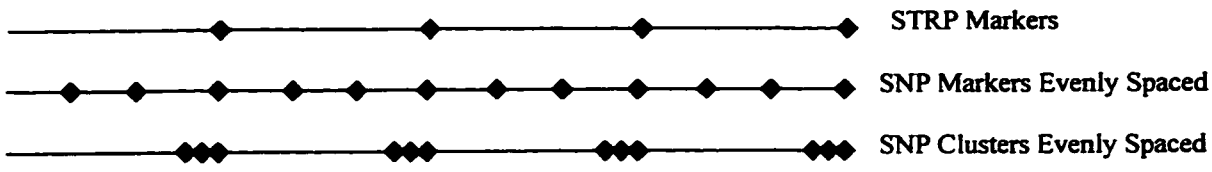


Figure 3.1 Possible genetic map structures.

In the first map structure, STRP markers (◆) are evenly spaced on the chromosome. In the second map structure, SNP markers (◆) are evenly spaced on the chromosome. Finally, in the third map structure, the clusters of markers (◆◆◆) are evenly spaced on the chromosome.

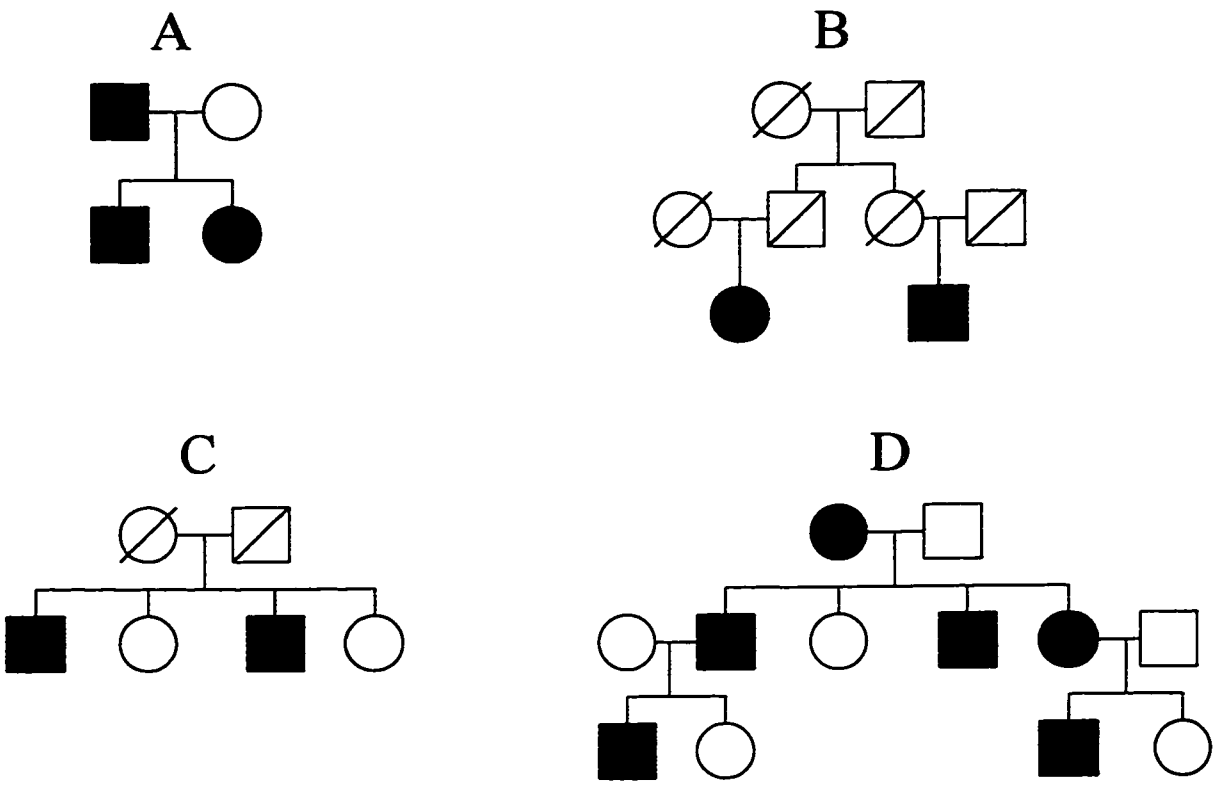


Figure 3.2 Pedigree structures used in the comparison of genetic maps.

A. affected sib-pairs B. affected cousin-pairs, C. nuclear pedigree, and D. extended pedigree. Individuals with a slash (/) did not have phenotype data, but marker data was included for all individuals.

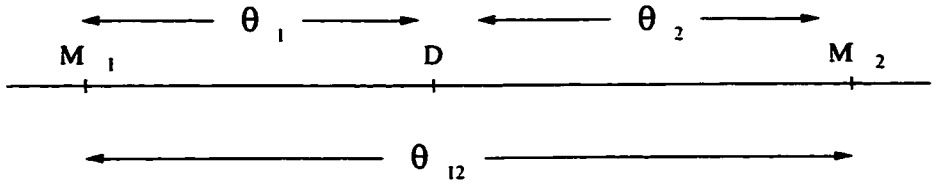


Figure 3.3 Genetic map for considering map inaccuracy.

Map showing the relationship between the two marker loci (M_1 and M_2), and the disease locus (D) in the case of linkage. The recombination fractions between the loci are denoted by θ_1, θ_2, and θ_{12}.

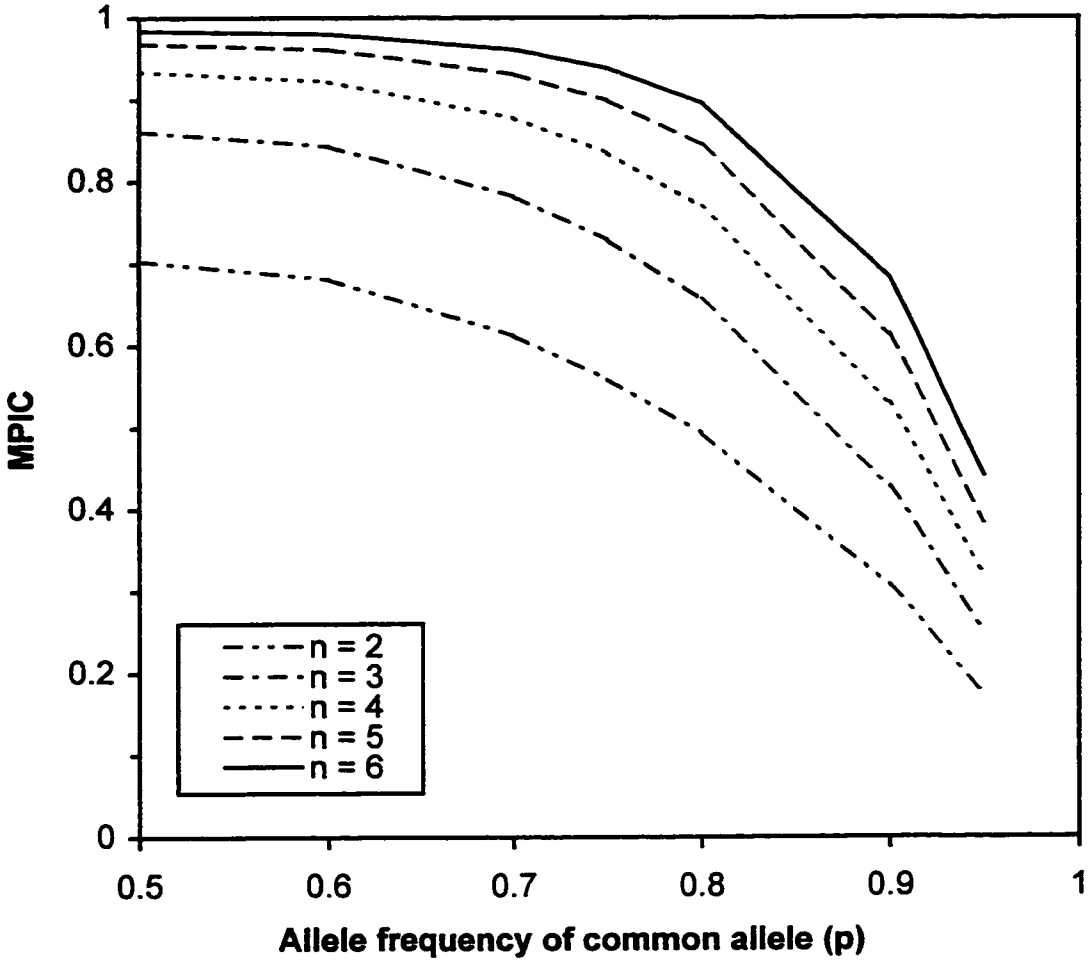


Figure 3.4. Effect of the allele frequency on MPIC. n: number of markers per cluster, $\pi = 1.0$, $\delta = 0$.

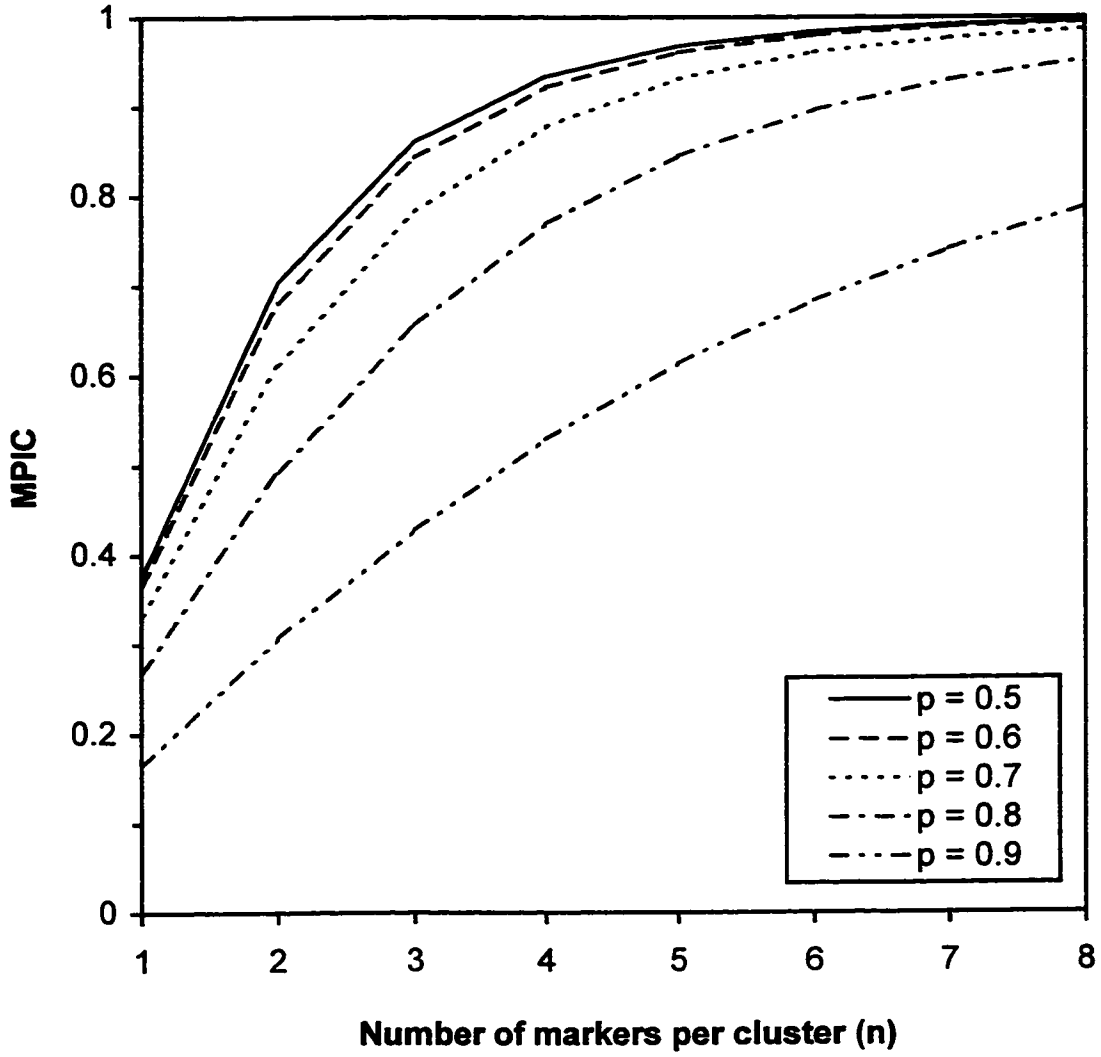


Figure 3.5. Effect of the number of markers per cluster on MPIC. p: allele frequency of common allele, $\pi = 1.0$, $\delta = 0$.

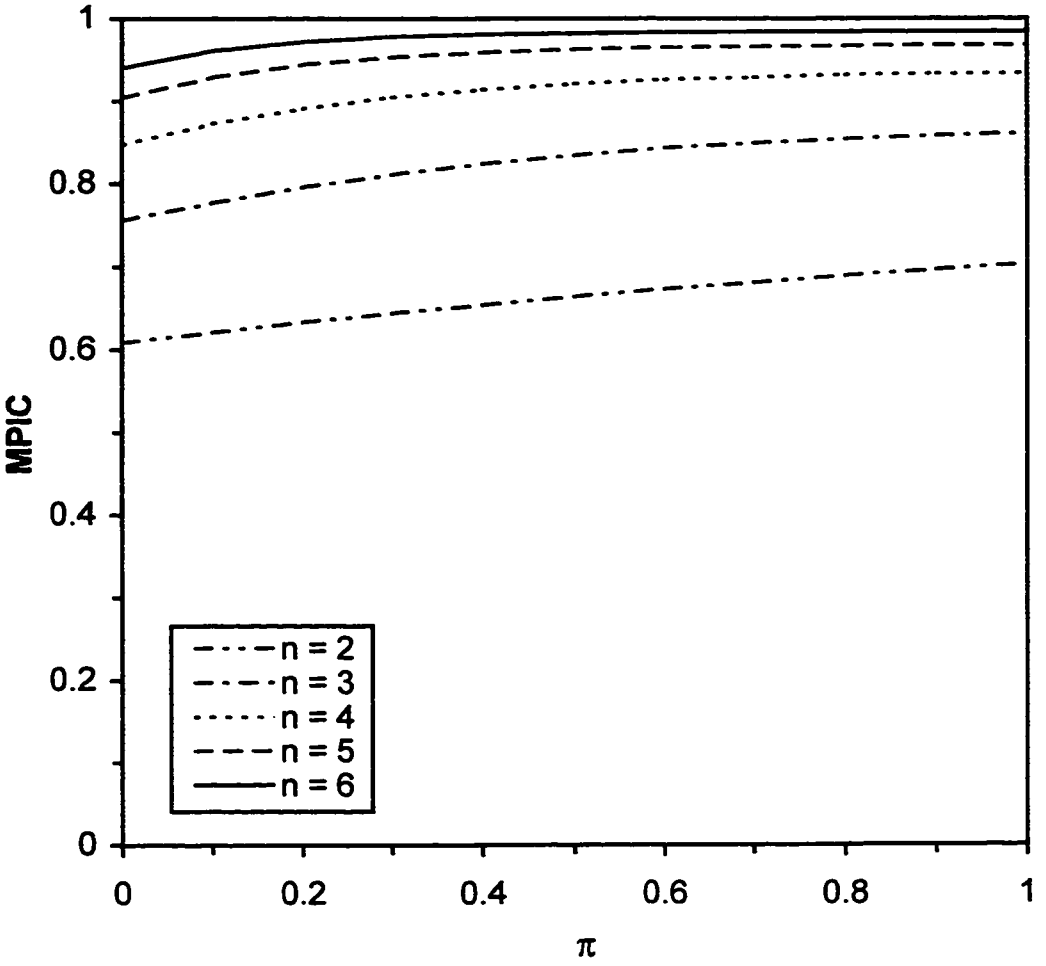


Figure 3.6. Effect of the probability of phase information on MPIC. n: number of markers per cluster, π: probability of phase information, $p = 0.5$, $\delta = 0$.

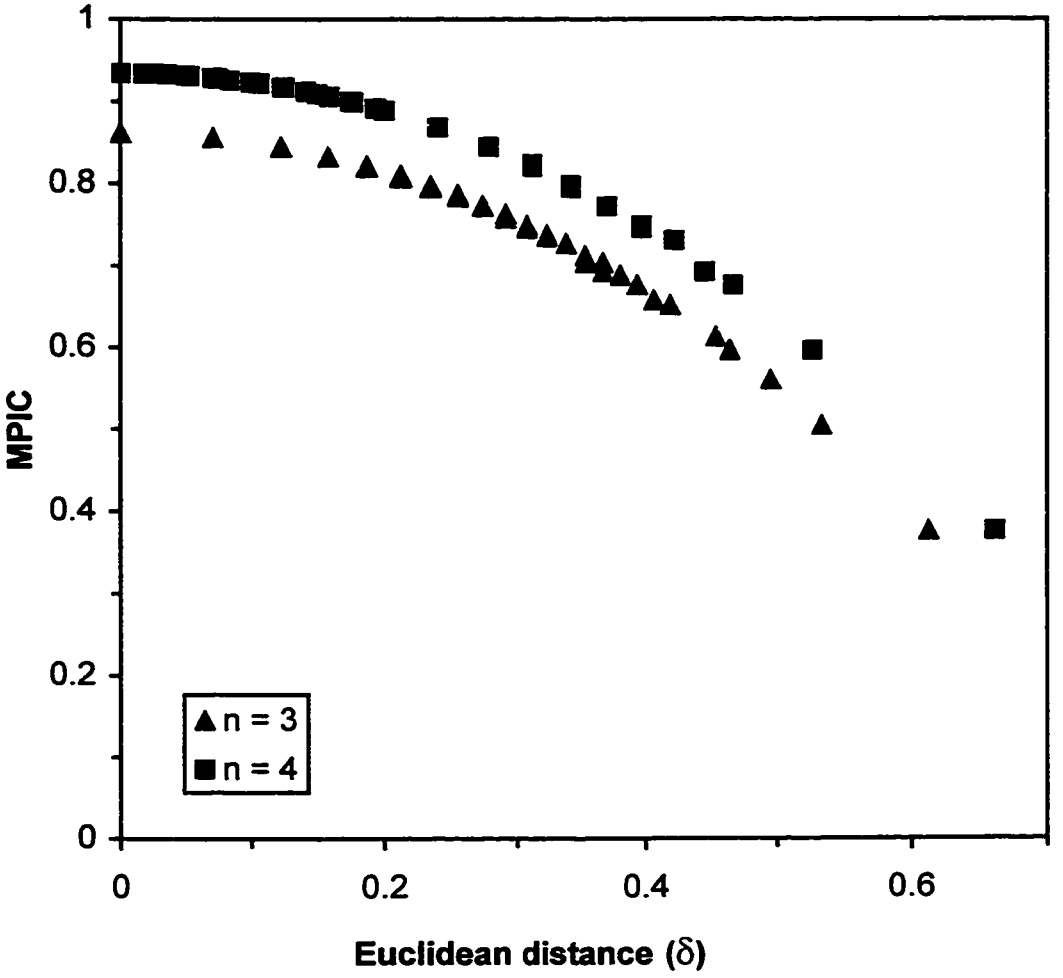


Figure 3.7. Effect of linkage disequilibrium on MPIC. n: number of markers per cluster, p = 0.5 for all loci in each cluster, $\pi = 1.0$.

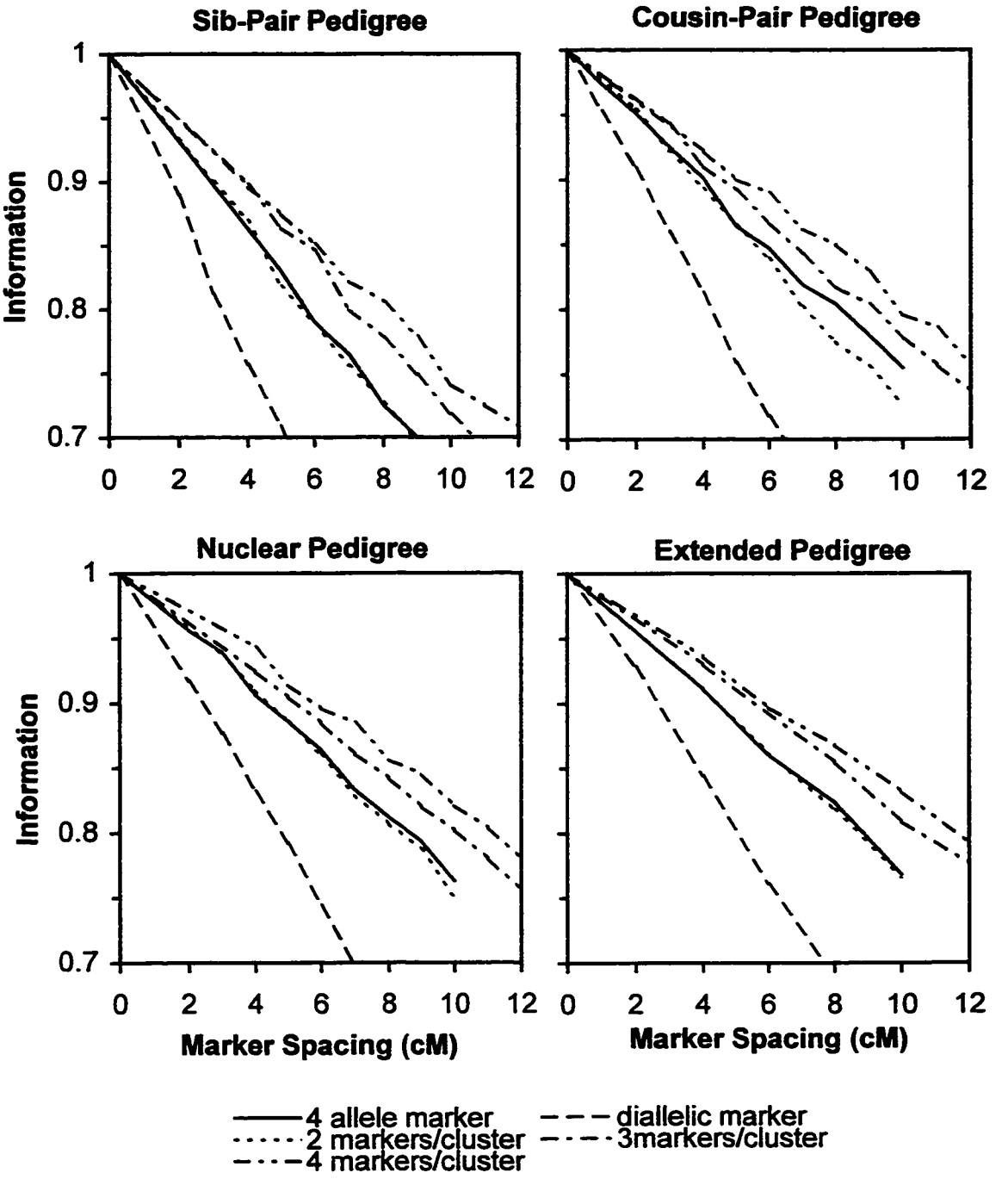


Figure 3.8. The effect of marker spacing on the information available from the genetic map. Five map structures are compared for each pedigree type. $\delta = 0$, $\pi = 1.0$, $p = 0.5$ for all loci in the cluster.

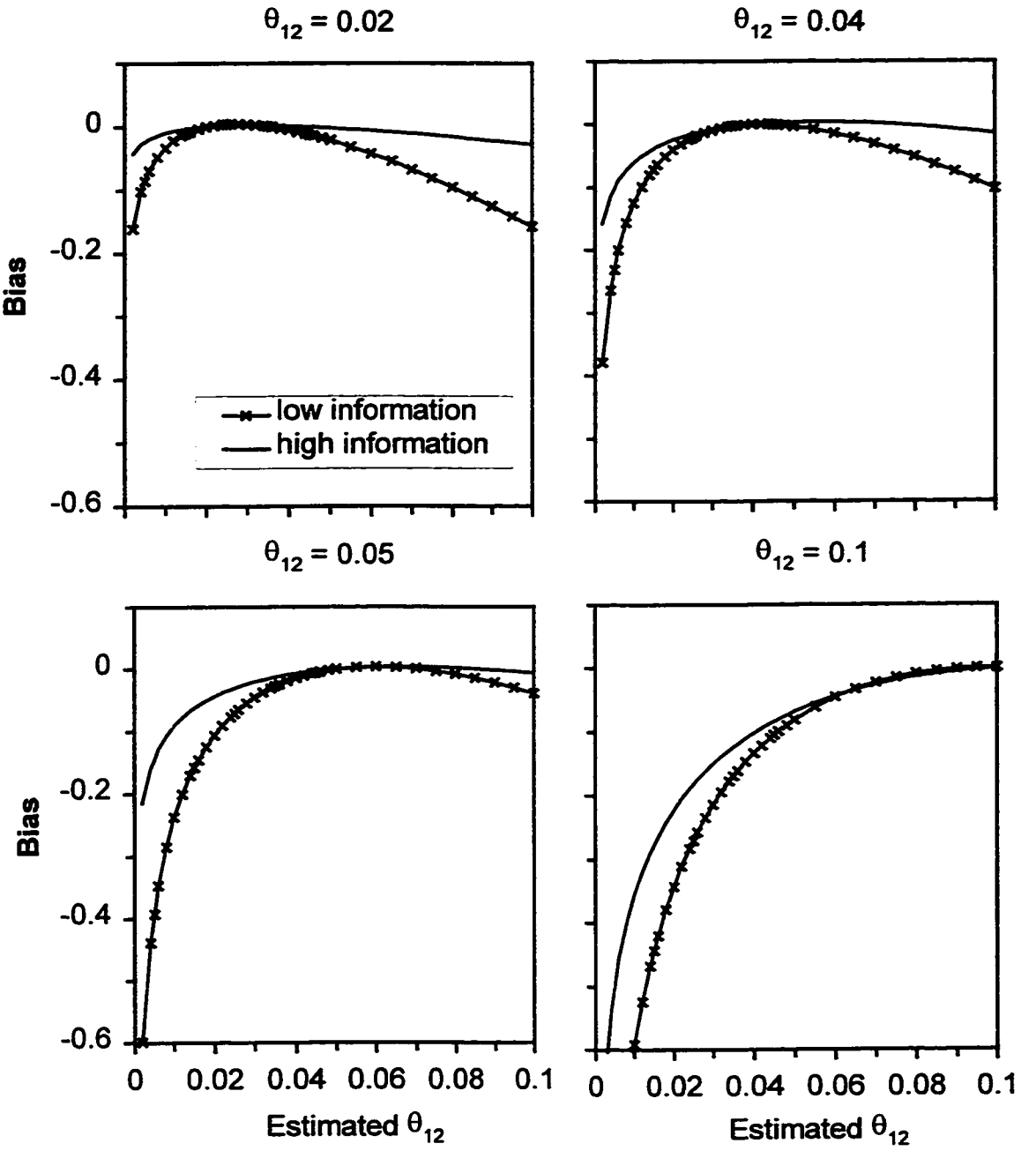


Figure 3.9. The bias from misspecification of the recombination fraction when the disease locus is located halfway between the marker loci. θ_{12} = recombination fraction between markers. The low information marker has two alleles, and the high information marker has 20 alleles.

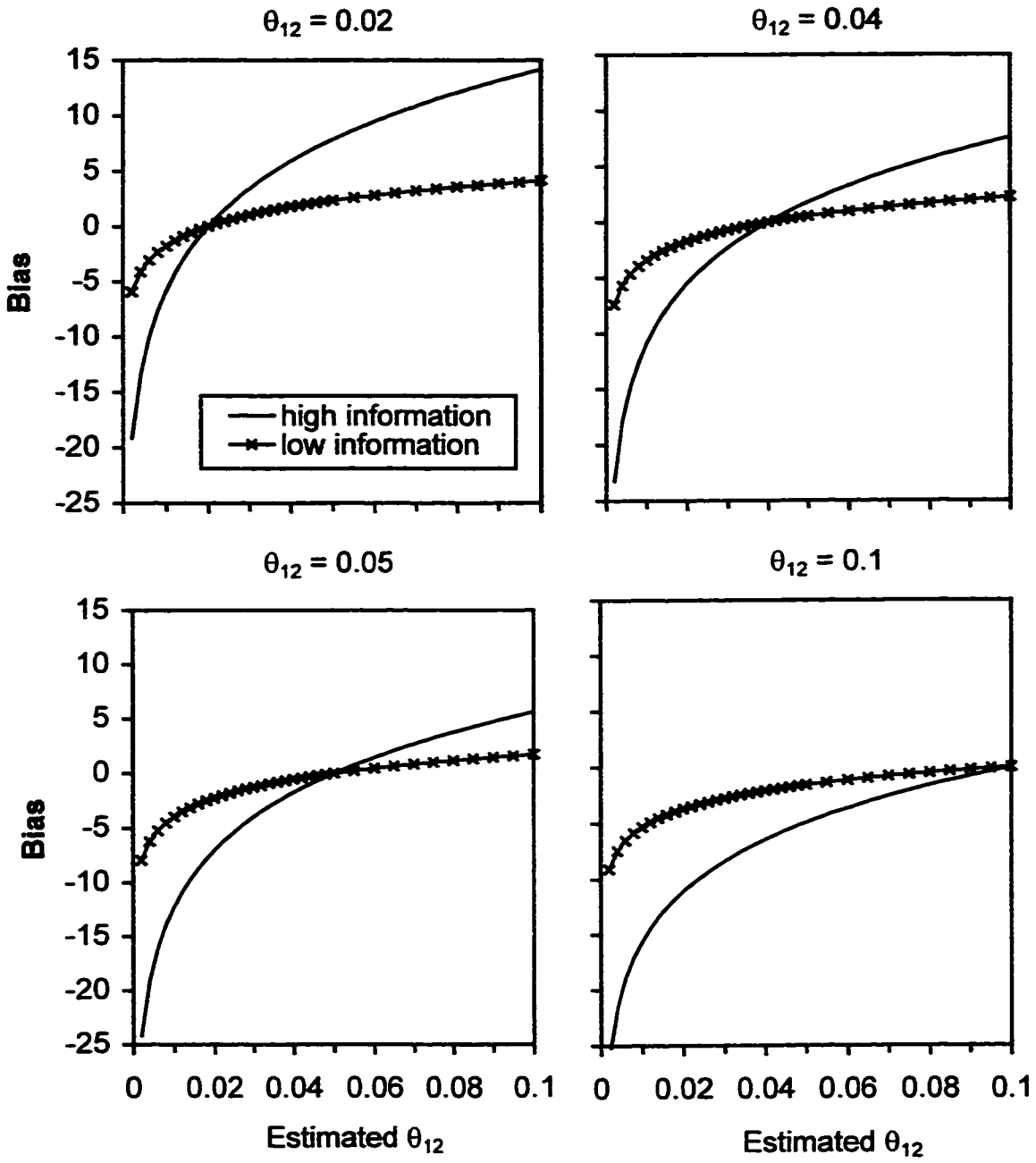


Figure 3.10. The bias from misspecification of the recombination fraction when the disease locus is unlinked to the marker loci. θ_{12} = recombination fraction between markers. The low information locus has 2 alleles, and the high information locus has 20 alleles.

Table 3.1 Possible genotype configurations for two parents and an offspring. Categories C1, C2, and C3 are always informative. Category C8 is never informative, even when combined with other loci. Categories C4 – C7 are sometimes informative depending on phase information. The notation homo (i) indicates that the individual is homozygous for allele i (i = 1,2). homo = homozygous, hetero = heterozygous.

	Class	Parent 1	Parent 2	Offspring
Always Informative	C1	hetero	homo	homo
	C2	hetero	homo	hetero
	C3	hetero	hetero	homo
Sometimes Informative	C4	hetero	hetero	hetero
	C5	homo	hetero	hetero
	C6	homo	hetero	homo
	C7	homo (i)	homo (j)	hetero*
Never Informative	C8	homo (i)	homo (i)	homo (i)*

*this is the only possible genotype for the offspring given the parental genotypes.

Table 3.2 Ratio of the number of markers necessary for the given map designs compared to a 10 cM screen using STRP markers.

Model	Sib-pairs	Cousin-pairs	Nuclear Pedigree	Extended Pedigree
Uniform SNP Design	1.6	1.8	1.8	1.7
2 SNPs per cluster	2.0	2.2	2.1	2.0
3 SNPs per cluster	2.5	2.7	2.6	2.5
4 SNPs per cluster	3.0	3.2	3.0	3.0

Table 3.3 Average bias in the Emlod from underestimating the marker distances.

n: number of informative meioses, θ_{12}: true recombination fraction between the markers.

	Low Information Content (2 alleles)		High Information Content (20 alleles)
θ_{12}	n = 28	n = 200	n = 200
0.02	-.161	-.037	-.010
0.04	-.176	-.021	-.014
0.05	-.257	-.030	-.013
0.10	-.209	-.012	-.014

CHAPTER 4 : DOWNCODING

4.1 INTRODUCTION

Likelihood-based linkage analysis approaches such as two-point and multipoint lod score methods are often more powerful than alternative methods for detecting linkage. Unfortunately, we are sometimes limited in our ability to utilize these powerful analysis methods because of computational constraints within the available resources. The two main limiting factors are the computation time and the memory requirements. In general, in using these approaches we typically maximize the likelihood over one or more of the model parameters. If there is incomplete data available, we must sum over all possible values of the missing information to calculate the likelihood. This is the source of the large computational burden since the computation time increases with the number of possibilities, and there must be enough memory to consider all of the possibilities.

Samples that are collected for studies of complex traits often have characteristics that increase the computational burden, including missing data, complex models, and large pedigree structures. Missing data such as missing marker genotype or trait phenotype information, reduced penetrance models, or phase ambiguity between the marker and the trait loci increases the number of possibilities that must be considered in computing the likelihood. Complex models generally include more parameters, which can also increase the number of possibilities that must be considered. Finally, large pedigrees are sometimes collected for studies of complex traits to attempt to reduce the heterogeneity present in the sample. The computation time is increased for large pedigrees since the joint probability of all of the individuals in the pedigree must be considered. It is important to identify methods to reduce the computational burden associated with the more powerful methods to allow their use in analyses of complex traits.

The number of alleles at the marker locus (or loci) is one additional factor that can increase the computational burden. Therefore, efforts to reduce the number of marker alleles that must be considered in the computations can have a significant impact on the feasibility of an analysis. Markers provide information about the transmission pattern of the chromosomal region in the pedigree. If the marker alleles are transmitted with the disease phenotype, there is evidence for a gene that contributes to the trait in the region. The information available about the transmission of marker alleles generally increases with the number of alleles at the marker. However, for most algorithms, the computation time and memory requirements also increase with the number of marker alleles. Selecting markers with fewer alleles is one method that could be used to reduce the computational burden. However, it may be difficult to obtain the additional pedigrees or additional markers that are necessary to compensate for the loss in information. Our goal is to retain the information about transmission in the pedigree that is provided by the more polymorphic markers without increasing the computational burden. In many instances, we can reduce the number of alleles that are used without loss of information about the transmission pattern of the marker. This process of allele reduction is called downcoding.

Algorithms for downcoding have been described previously. However, each method is limited in the type of pedigree structure that can be used or in the efficiency of the reduction. Ott's downcoding method (Ott, 1978) is limited to pedigrees with simple structure (no inbreeding loops or multiple founder sets) and nearly complete data. Unfortunately, the situations where downcoding is most needed are pedigrees with complex structure or missing data, which do not fit Ott's criteria. Braverman's downcoding method (Braverman, 1985) allows for complex structure pedigrees and missing data, but does not always reduce the number of alleles as efficiently as possible.

Several alternative methods have been proposed in addition to downcoding. One method is to reassign allele frequencies for each pedigree so that the number of alleles is reduced to the number observed in the pedigree plus one additional allele which represents all the remaining alleles (Schellenberg et al., 1992). Unfortunately, this method does not affect

the computation time if there is only a single pedigree, and also does not always reduce the number of alleles as efficiently as possible. A second method, which is implemented in the program VITESSE, assigns alleles to sets to reduce the computation time (O'Connell and Weeks, 1995). In the current implementation, this method cannot be used for pedigrees with inbreeding loops or multiple founder sets. This is an important limitation since pedigrees with these structures generally have a higher computational burden than pedigrees with a simple structure. Finally, a hidden Markov model algorithm (Lander and Green, 1987) implemented in the program GENEHUNTER (Kruglyak et al., 1996) for multipoint or two-point linkage analysis allows a large number of markers to be evaluated simultaneously, and is not affected by the number of alleles at the marker loci. Unfortunately, this algorithm is limited to evaluation of small or moderate size pedigrees, because of the memory requirements for larger pedigrees.

Here we describe a method of downcoding that can be used for arbitrary structure pedigrees and with missing marker data. Downcoding has been used by many researchers through performing the downcoding by hand (e.g. van Duijn et al., 1994). This is time consuming and error prone, especially on large pedigrees where the gains are most apparent. The development of an algorithm to perform downcoding can define the principles underlying the process and can also be useful for developing algorithms for pedigree computations in general.

4.2 METHODS

To begin, we make some assumptions about the structure of the data. We assume that each individual in the pedigree has two alleles that can be observed for a particular marker locus, and that either both alleles are known, or neither allele is known initially. An individual is a founder if his or her parents are not included in the pedigree; otherwise, the individual is a non-founder. We assume that both parents are included in the pedigree for all non-founders.

If there are multiple marker loci that are used in the analysis, we downcode each marker locus individually without considering the alleles that are observed at the other marker loci. We will show later that results obtained this way are valid for multipoint analysis. The allele labels are the names that are used to identify different alleles. We assume that there are n allele labels observed in the pedigree, and our goal is to reduce the number of allele labels that must be considered in the analysis.

The method of downcoding presented here proceeds by assigning all alleles observed in the pedigree to sets that must have the same allele label to retain the transmission information, and then reusing allele labels for different sets to reduce the total number of allele labels in the pedigree. There are three main steps to this process. First, alleles that are identical by descent (IBD) must have the same allele label, so step 1 identifies sets of alleles that are known to be IBD. Second, a pair of alleles must also retain the same label if we cannot eliminate the possibility that they are IBD. Therefore, step 2 combines sets that must retain the same allele label for this reason. Finally, allele labels are assigned to the sets so that the transmission information is maintained, but allele labels are reused when possible to reduce the number of alleles observed in the pedigree. We describe rules for performing the relabeling of allele sets in step 3.

4.2.1 Step 1

All alleles that are identical by descent (IBD) must have the same allele label. We consider each non-founder with marker genotype information along with his or her parents to determine whether we can identify the IBD status of the alleles. For convenience in this discussion, we refer to the offspring as the index case, and to the parents as the mother and the father. There are three possible conclusions from the information available from the three individuals. First, a determination can be made about the IBD status of the alleles observed in the index case. It is not possible to determine the IBD status of only one allele. For example, if a particular allele is transmitted from the mother to the offspring, then the other allele must be transmitted from the father. Second, we may conclude that

there is an incompatibility. In this case, it is impossible for the index case to have the observed genotype given the information available about the parental genotypes, and the algorithm will terminate. Finally, we may conclude that the IBD status of the alleles observed in the index case is ambiguous. This case is discussed later, where we show how other pedigree members may provide sufficient information to determine the IBD status.

We can determine the IBD status of the alleles in the index case if it is impossible for one of the alleles to have been transmitted from one of the parents. For example, if the father has genotype {A,B}, and the index case has genotype {A,C}, then the C allele could not have been transmitted from the father, and we know that the A allele observed in the index case is IBD to the A allele observed in the father. Since we cannot know the IBD status of one allele without knowing that of the other, the C allele observed in the index case must be transmitted from the mother unless there is an incompatibility.

There are three situations where the IBD status of alleles may be ambiguous. First, if all three individuals have the same heterozygous genotype, then both alleles that are observed in the index case could have been transmitted from either parent, and the IBD status is unclear. Second, if one or both parents are homozygous for the marker genotype, then we cannot determine which allele is transmitted to the index case because both alleles have the same allele label. In both of these cases, there is no additional information that could be provided to resolve the ambiguity. Third, missing data in one or both of the parents can also be a cause of ambiguity. In this case, we may be able to identify a set of possible alleles for the individual(s) with missing data given information about the genotypes of other pedigree members. We do not need to identify the genotype of the individual(s) with missing data, we only need to identify a complete set of alleles that the individual *could* have transmitted.

There are three sources of information that could resolve an ambiguous situation due to missing data including siblings of the index case (both half-sibs and full sibs), ancestors of the parent(s) with missing information, and siblings of the parent(s) with missing in-

formation. In each case it may be possible to determine a set of possible alleles that could be transmitted by the parent(s) with missing data from the information provided by the other relatives in the pedigree. We can then compare the alleles in this set to the alleles observed in the index case to potentially determine the IBD status of the alleles. These situations can be illustrated with a few examples.

A set of possible alleles that could be transmitted by an individual with missing data can be defined by the alleles observed among the offspring of this individual. Consider a nuclear pedigree where the father has genotype {A,B} and the mother has unknown marker genotype. The index case also has genotype {A,B}, so we cannot determine which allele was transmitted from the father. If we observe alleles {A,B,C,A} among all of the offspring of these two parents (i.e. one of the offspring has genotype {A,A}), then it is clear that the mother has genotype {A,C}, so the index case must have received the A allele from the mother, and the B allele from the father.

In addition, the set of possible alleles can be defined by the alleles observed among any collection of ancestors with no missing data such that any line of descent from a founder to the individual passes through exactly one ancestor in the collection. For example, consider a pedigree with the same initial conditions for the genotypes of the two parents and the index case as in the previous example. In this example, the genotypes of the maternal grandparents are known to be {A,C} and {A,D}. Although we do not know the genotype of the mother, we do know that the only possible alleles the mother could have transmitted are {A,C,D}. Therefore, the index case must have received the A allele from the mother, and the B allele from the father.

Finally, the set of possible alleles can be defined by the alleles observed among the siblings of the individual with missing data. Beginning with the same initial conditions as before, if the alleles observed among the siblings of the mother are {A,C,D,E}, then although we do not know what the maternal grandparental genotypes are, we know that the

mother could not have transmitted a B allele. Therefore, the index case must have received an A allele from the mother, and a B allele from the father.

In cases where there is ambiguity in the IBD status for individual because of missing data in the parents, we may be able to use information from other members of the pedigree to resolve the ambiguity. Note that we must be able to identify a complete set of possible alleles for the parent with missing data. For example, if we only know three of the four alleles among the maternal grandparents in the example given above (i.e. {A,C,D,?}), then it may still be possible for the mother to have transmitted the B allele.

4.2.2 STEP 2

All alleles where we cannot eliminate the possibility that the alleles are IBD must have the same allele label. If two alleles have different initial allele labels, then it is not possible that they are IBD, so we only need to consider cases where the alleles have the same initial allele label. A set of alleles is composed of all of the alleles that must have the same allele label. The rules described below determine the assignment of alleles to sets.

1. All alleles that are IBD are assigned to the same set.
2. For alleles where an ambiguity could not be resolved in step 1, find the most recent direct ancestor in each line of descent with an allele with the same allele label. Assign these two alleles to the same set.
3. Consider two individuals with alleles with the same allele label, and where an ambiguity could not be resolved in step 1 for both individuals. If the individuals have a common ancestor, and there are no individuals connecting the two individuals to their common ancestor that are known to have an allele with this allele label, then assign these two alleles to the same set.
4. If there is an allele observed in a founder that is not transmitted to any offspring, then assign this allele to its own set.

In step 1, we described three situations where there might be an ambiguity in the IBD status of alleles for the index case. The first two cases where one of the parents is homozygous, or where all three individuals are heterozygous for both alleles can be resolved by rule two. In each case, the alleles observed in the index case maintain the same allele labels as the alleles in either parent that could potentially be IBD. In the remaining cases where there is missing data, both rules two and three apply. Here, an example may best illustrate these cases. For the pedigree in figure 4.1, allele A observed in individual 10, and allele A observed in individual 1 must be assigned to the same set because of rule two. Individual 1 is the most recent direct ancestor of individual 10 with an A allele. Also in this pedigree, allele D observed in individual 11 and allele D observed in individual 8 must be assigned to the same set because of rule three. These two individuals have a common ancestor (either individual 3 or 4) who could have transmitted the D allele to both individuals, and there are no individuals connecting individuals 11 and 8 with a D allele (i.e. individuals 7, 3, or 4).

4.2.3 STEP 3

The final step is to relabel the allele sets to reduce the total number of allele labels in the pedigree. In general, there are two possible ways to reduce the number of alleles observed in a pedigree. First, we may be able to reduce the number of alleles that are required for a particular mating while still retaining the transmission information. For example, for the pedigree in figure 4.2, there are four alleles observed in the parental genotypes. We can relabel the D allele to a B without losing the underlying transmission information, leaving only three alleles present in the pedigree. Second, we may be able to maintain the same number of alleles among the parents, but use different allele labels for one or more of the observed alleles to eliminate allele labels that are not observed elsewhere in the pedigree. In the pedigree in figure 4.3, there are three alleles observed in the mating in the middle generation which are A, B, and E. By relabeling allele E to allele C, there are still three alleles observed in this mating, but we have reduced the number of alleles observed in the

pedigree from five {A,B,C,D,E} to four {A,B,C,D}. The number of alleles observed in the mating in the top generation can be reduced from four to three, so we can reduce the total number of alleles observed in this pedigree to three {A,B,C}.

More specifically, the relabeling proceeds as follows. All of the alleles are assigned to sets denoted S_1, ..., S_n in steps 1 and 2, where the maximum value of n is two times the number of founders in the pedigree. The correct allele frequency must be used in the likelihood calculation for alleles with an ambiguity in the assignment of the IBD status due to missing information in the parent(s) (see appendix 2). Therefore, the minimum number of allele labels necessary to relabel the pedigree depends on the number of sets that contain alleles with this type of ambiguity. We refer to the new allele labels as L_1, ..., L_t, where there are $0 \leq t \leq n$ sets initially that must have a distinct label. If there are no sets that must have a distinct label because of ambiguity in the IBD status of alleles, then we assign set S_1 to label L_1. It may be possible to reuse some of these allele labels for the remaining sets, so L_i corresponds to a vector of allele sets that are assigned to allele label i, where L_{ij} refers to the jth component of the vector.

The remaining sets are relabeled by considering the configuration of alleles in each parent-offspring trio. In Appendix I all possible configurations of genotypes for two parents and an offspring are considered and corresponding rules for each case are described. For each parent-offspring trio, we determine whether the observed allele sets can have the same new label. There are three possible outcomes for the comparison of sets S_i and S_j: if $\{i,j\} = 0$ the sets cannot be assigned the same new allele label, if $\{i,j\} = 1$ the sets can be assigned the same new allele label, and if $\{i,j\} = 2$ the sets can be assigned the same new allele label *if* the other two sets that are observed in the parents, S_k and S_l, are not assigned the same new allele label. Therefore, if $\{i,j\} = 2$, we have the condition, C_{ij} = {if S_k is assigned to L_a, and S_l is assigned to L_b, then $a \neq b$}. The default state for all pairs of allele sets is $\{i,j\} = 1$. We proceed through each observed parent-offspring trio, and change the value of $\{i,j\}$ to 0 or 2 if necessary according to the rules described in the ap-

pendix. If the value of $\{i,j\} = 0$ for a particular parent-offspring trio, it must remain 0, and if the value of $\{i,j\} = 2$, it cannot be changed back to 1, although it can be changed to 0 from consideration of additional parent-offspring trios. There can be multiple conditions for comparisons with outcome 2 from different matings that are observed in the pedigree.

Based on the information from the pairwise comparisons, we assign the remaining sets to a new allele label only if it is not possible to assign the set to one of the existing allele labels. It is not possible to assign set S_i to one of the existing allele labels if one of the following cases is true for at least one j in each group L_k, $k = 1, \ldots, t$, 1) $\{i,j\} = 0$ or 2) $\{i,j\} = 2$ and C_{ij} is true, or 3) $\{i,j\} = 1$ makes C_{ab} true for any sets a and b that have already been assigned. In this case, assign set S_i to a new allele label, L_{t+1}. Otherwise, there exists an allele label, L_k, where it is consistent to assign set S_i to label L_k. We proceed through all of the remaining allele sets until all sets have been assigned a new label.

We do not lose allele transmission information from the relabeling procedure since the rules for relabeling allele sets are based on considering the parent-offspring trios. This relabeling procedure will work for any structure pedigree because we are simultaneously considering all of the matings to determine how to relabel the allele sets. Therefore, the rules for assigning new allele labels will maintain the transmission information for all of the observed matings in the pedigree.

4.3 RESULTS

The method of downcoding presented here is more efficient than previous methods in reducing the number of alleles. Figure 4.4 depicts an example in which six alleles are observed in the original pedigree. Braverman's downcoding method reduces the number of alleles from six alleles to four, while the method presented here reduces the number of alleles to three. Braverman's method is inefficient because it does not allow alleles that occur in the same mating to be reassigned to the same allele label in some circumstances where it should be possible. However, we are able to make such allele label reassignments for the mating in the middle generation without loss of transmission information.

In a second example (Figure 4.5), Braverman's downcoding method does not reduce the number of alleles, while the method presented here reduces the number of alleles by two (Figure 4.5). Braverman's method does not reduce the number of alleles within nuclear families that contain at least one individual with missing data, although allele reduction can occur in other portions of a more extended pedigree. As we demonstrate in this example, downcoding can reduce the number of alleles for nuclear families with missing data without loss of transmission information.

The computation time for two-point linkage analysis increases with the number of alleles observed in the pedigree. Figure 4.6 shows the computation time as a function of the number of alleles for the pedigree structure in figure 4.7. This example illustrates a problem where downcoding may be required for two-point linkage analysis. One analysis with 10 alleles required more than 6 days of computation time for this pedigree. We were able to downcode this marker to 8 alleles, which reduced the computation time to approximately 28 hours without changing the lod scores. Additional examples of the reduction in computation time from downcoding are listed in table 4.1. Alternative methods such as the algorithms implemented in VITESSE or GENEHUNTER are not available for this pedigree due to the large number of individuals in the pedigree, and the complex structure including an inbreeding loop and multiple founder sets.

The computation time for multipoint linkage analysis also increases with the number of alleles at each locus observed in the pedigree. We demonstrate the reduction in computation time that can be achieved through downcoding for two multipoint examples (Table 4.2). In these cases, the computation time for the original data is 20 to 100 times the computation time required for the downcoded data. As expected, the lod scores for the multipoint analyses were not affected by downcoding (Table 4.2). The VITESSE algorithm can perform the computations for either multipoint analysis using the original data in less than five seconds, which is significantly faster than the computation time of 48 hours for the downcoded data in the second example . The computation time is not significantly affected by combining downcoding with VITESSE for these examples. It is not

possible to use the algorithm in GENEHUNTER for these examples because of the number of individuals in the pedigree.

4.4 CONCLUSIONS

We have presented a new method of downcoding that can be used for pedigrees with missing marker genotypes and with complex structure. This method is more efficient than previously published methods in reducing the number of alleles because it allows downcoding in portions of the pedigree that contain missing information, and it allows alleles within a mating to be relabeled to the same allele label. In addition, this method of downcoding can be used for a pedigree with arbitrary size and structure. Many of the other methods that are currently available for reducing the computation time for linkage analysis are restricted in the number of individuals, or the structure of the pedigree that can be evaluated.

Downcoding should not be used in all cases. If linkage disequilibrium is present among loci in the analysis, downcoding with the current algorithm should not be used. In this case, information about the alleles present at one locus may provide information about phase or missing marker genotypes at another locus. In addition, downcoding should not be used if we are interested in calculating the likelihood instead of the lod score or some other form of the likelihood ratio. As we show in appendix 2, downcoding can alter the likelihood, although the changes from downcoding do not alter the likelihood ratio. Finally, downcoding should not be used in situations where we allow for error in the marker genotypes, since the probability of the marker phenotype given the genotype will depend on the exact alleles that are observed. Usually we assume that the marker alleles are accurately known, but there may be situations where we would want to allow for error in the genotyping since this process is known to produce errors (approximately 0.5% error rate in many large laboratories).

We have demonstrated that the computational burden associated with several potentially powerful linkage analysis approaches can be reduced through the use of downcoding. For complex traits that often require large sample sizes, the use of efficient analysis methods can have a significant impact on the overall cost of a study by reducing the necessary quantities of other resources such as pedigree collection or marker genotyping that are required to detect linkage.

4.5 APPENDIX 1

To determine how to relabel the allele sets, we only need to consider the transmission pattern of the alleles in each nuclear family. If downcoding does not change the transmission information in each nuclear pedigree, then the information for the whole pedigree will also not be affected. Each allele has an initial label and a final label. The initial label is the allele label that was in the data file. After downcoding, each allele will have a final label that may be different from the initial label. The final label for two alleles that are in the same set will be the same by definition. The final label for two alleles in different sets may be the same or different. The procedure for relabeling sets is to compare each allele observed in the parental genotypes (a total of six comparisons) to determine if the alleles can have the same final label. There are three possible outcomes for the comparison of allele *i* in set S_i and allele *j* in set S_j:

- Outcome 0: S_i and S_j cannot have the same final label,
- Outcome 1: S_i and S_j can have the same final label, and
- Outcome 2: S_i and S_j may be able to have the same final label *if* the other two alleles that are observed in the mating do not have the same final label.

To determine which outcome will occur for a particular comparison, rules have been defined based on the "mating type" of the individuals under consideration. The mating type refers to the pattern of known and missing data in the individuals without regard to the particular allele labels that are observed in the individuals. The mating type is defined for a trio of individuals that include the two parents and an offspring. An example of the no-

tation of a mating type is {P1: XX, P2: X?, O: ??} which indicates that parent 1 has two alleles observed, parent 2 has one allele observed and one allele missing, and the offspring has both alleles missing. The X's do not indicate a particular allele label (for instance, parent 1 is not necessarily homozygous), they only indicate whether or not the allele is missing. The notation {i,j}=1 will be used to indicate that the comparison between allele i and allele j had outcome 1 with similar notation for the other two outcomes.

The rules for assigning new labels that are listed here do not depend on the mating type.

1. If allele i and allele j ($j \neq i$) have the same initial label, then {i,j}=1.
2. If either allele i or allele j is unknown, then {i,j}=0.
3. If allele i and allele j are both in the same parent, and the initial label for i is not the same as the initial label for j, then {i,j}=0; therefore, a heterozygous parent cannot be relabeled as homozygous.
4. If an offspring of the mating has genotype (ij), the offspring has descendants with marker data, and the IBD status is known for the alleles in the offspring, then {i,j}=0. This rule implies that if the offspring is at the bottom of the pedigree, or there is no information for any individuals who are descendants of the offspring, then the offspring can be relabeled to be homozygous without loss of information about transmission in the pedigree.

There are some rules for assigning new labels that do depend on the mating type, which are listed here. However, there is not a specific rule for every mating type, so the following is not a complete list of all mating types. Relabeling for mating types that are not listed here must still be consistent with the rules listed above, but otherwise the alleles in these mating types are not downcoded.

CASE 1. (P1: XX, P2: XX, O: X?)
If the allele observed in O is in set S_i, then for any allele observed in the parents that is in set $S_j \neq S_i$, {i,j}=0.

CASE 2. (P1: any,P2: any,O: ??)

If O does not have any descendants with marker data, then unless stated otherwise in rules one through four given above, {i,j}=1 for all comparisons. If O does have descendants with marker data, then unless stated otherwise in rules one through four given above, {i,j}=0 for all comparisons. If allele k is observed in a descendant of O, and allele i is an allele observed in this mating, if $k \neq i$ for all alleles observed in this mating, then {i,k} = 0 for all values of i.

CASE 3. (P1: XX, P2: X?, O: XX)
If allele i is observed in one parent, and allele j is observed in the other parent, then {i,j}=0 unless rule one given above applies.

CASE 4. (P1: XX, P2: X?, O: X?)
(P1: X?, P2: X?, O: X?)
(P1: X?, P2: ??, O: X?)
(P1: X?, P2: ??, O: XX)
If allele i and allele j do not have the same initial label, then {i,j}=0. If allele k is observed in a descendant of O, and allele i is an allele observed in this mating, if $k \neq i$ for all alleles observed in this mating, then {i,k} = 0 for all values of i.

CASE 5. (P1: X?, P2: X?, O: XX)
If O is homozygous and allele i and allele j are known, then {i,j}=1. If O is heterozygous then {i,j}=0 unless rule one given above applies. If allele k is observed in a descendant of O, and allele i is an allele observed in this mating, if $k \neq i$ for all alleles observed in this mating, then {i,k} = 0 for all values of i.

CASE 6. (P1: XX, P2: XX, O: XX)
It is possible to relabel the parents to be heterozygous for the same two alleles if none of the offspring of the mating have descendants with data, and if all offspring of the mating can be relabeled as homozygous in such a way that it is not ambiguous which alleles were transmitted (i.e. all offspring have genotypes {A,C} or {B,D} given the parental genotypes of {A,B} and {C,D}). In this case, {i,j}=1 unless rules one through four given above apply. Otherwise, {i,j}=0 unless rules one through four given above apply.

Once all the other rules have been taken into account, any remaining comparison has {i,j}=2. Outcome 2 allows sets S_i and S_j to be relabeled to the same allele label if the other two alleles observed in the mating do not have the same allele label. This generally allows the reduction of alleles in a mating from 4 allele labels to 3 allele labels, while preventing a reduction to 2 allele labels when such a reduction would result in ambiguous transmission information.

Mating types that do not have a specific rule:

(P1: XX, P2: X?, O: X?)
(P1: X?, P2: X?, O: X?)
(P1: XX, P2: ??, O: X?)
(P1: ??, P2: ??, O: X?)
(P1: XX, P2: ??, O: XX)
(P1: ??, P2: ??, O: XX)

4.6 APPENDIX 2

The likelihood for a single pedigree can be written as:

$$L(\theta) = \Pi_i \Sigma_j P(Y_i|g_{ij}) P(g_{ij}|\cdot\cdot),$$

where θ is the recombination fraction, Y_i is the phenotype for individual *i*, and g_{ij} is the *j*th genotype for individual *i*. $P(g_{ij}|\cdot\cdot)$ is the probability of the genotype given the parental genotypes for non-founders, and the probability of the genotype in the population for founders. Following the notation of Ott (1978), for a pedigree with *n* individuals, we can rewrite the likelihood as:

$$L(\theta) = \sum_{i=1}^{n} \prod_1 P(Y_i|g_i) \prod_2 P(g_i) \prod_3 P(g_i|\cdot\cdot),$$

where the (multiple) sum is over all assignments of joint genotypes to each individual, the first product is over all individuals, the second product is over all founders, and the

third product is over all non-founders. The $P(Y_i|g_i)$ terms can be separated for each marker and trait locus, assuming linkage equilibrium between the loci such that if there are n trait and marker loci, $P(Y_i|g_i)=\prod_{j=1}^{n} P(Y_{ij}|g_{ij})$. For the trait locus (loci), the $P(Y_{ij}|g_{ij})$ term is not affected by the downcoding procedure since we do not downcode the trait locus. For the marker locus (loci), the $P(Y_{ij}|g_{ij})$ term is either 0 or 1 for individuals with known marker data since we assume the marker genotypes are correctly known, and this term is equal to 1 for individuals with unknown marker data. The $P(g_i)$ terms cancel out in the likelihood ratio to the extent that the founder marker genotypes are known because the $P(Y_{ij}|g_{ijk})$ is equal to 0 for individual i at locus j for all k where $Y_{ij} \neq g_{ijk}$. For example, if we know that founder individual 1 has marker phenotype AA, then all terms will have a^2 (if a is the allele frequency for allele A), which can be brought out in front of the summation. The likelihood becomes

$$L(\theta) = a^2 \sum_{i=2}^{n} \prod_1 P(Y_i|g_i) \prod_2 P(g_i) \prod_3 P(g_i|\cdot\cdot),$$

where the (multiple) sum is over the genotypes of the remaining individuals in the pedigree. In the likelihood ratio, the frequency of the known founder marker genotype is present in both the numerator and the denominator, so it cancels out.

$$LR = \frac{L(\theta_1)}{L(0.5)} = \frac{a^2 \sum_{i=2}^{n} \prod_1 P(Y_i|g_i) \prod_2 P(g_i) \prod_3 P(g_i|\cdot\cdot)}{a^2 \sum_{i=2}^{n} \prod_1 P(Y_i|g_i) \prod_2 P(g_i) \prod_3 P(g_i|\cdot\cdot)}$$

Correct allele frequencies are only necessary for founder genotypes that are ambiguous due to missing data since the second product is only over the founders. Therefore, in order for the likelihood ratio to be correct, all sets that contain alleles with ambiguous IBD status from missing data in the founders must be assigned a unique allele label that corresponds to the correct allele frequency. To show that downcoding does not affect the likelihood ratio, it suffices to show that the $P(g_i|\cdot\cdot)$ terms are not affected by downcoding for

the non-founders. To do this, we need to consider each possible parent-offspring trio to determine whether downcoding affects the probability of the observed offspring genotype given the parental genotypes. However, we have already considered these situations in appendix 1 when we stated the rules for assigning new allele labels. Therefore, if these rules are correct, the likelihood ratio should not be affected by the downcoding method that is presented here.

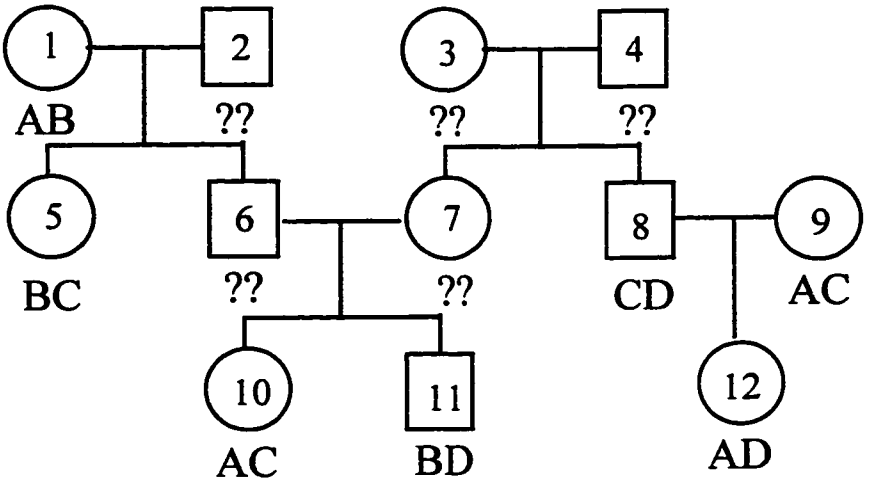


Figure 4.1 Example pedigree.

The individual ID number is shown within the symbol, and the genotype is below the symbol. If the genotype is ??, the individual has missing data.

Figure 4.2 Example showing downcoded pedigree.

The pedigree on the left is the original pedigree, while the pedigree on the right has been downcoded to reduce the number of alleles that are observed from four to three.

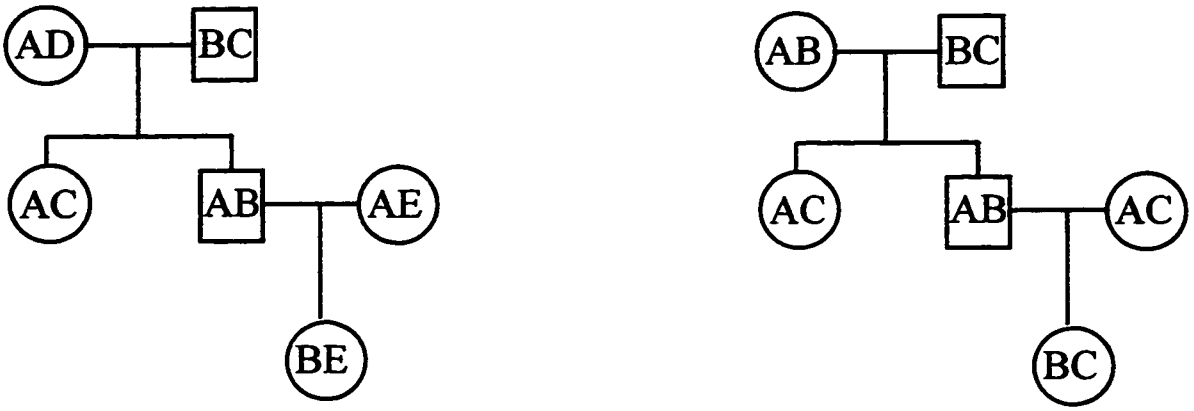


Figure 4.3 Example showing downcoded pedigree.

The pedigree on the left is the original pedigree, while the pedigree on the right has been downcoded to reduce the number of alleles from five to three.

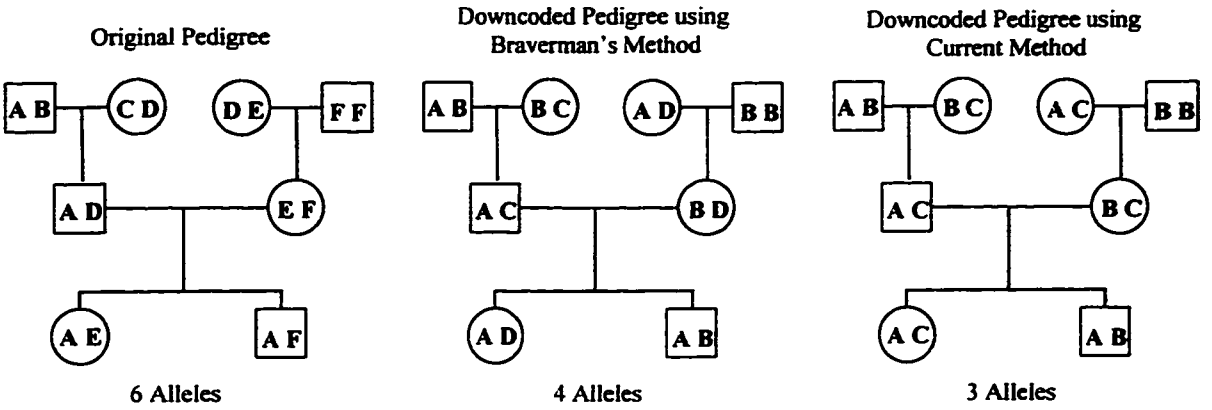


Figure 4.4 Example showing the reduction in the number of alleles from downcoding using two methods.

This example shows the reduction in the number of alleles for a completely typed pedigree using Braverman's method and the method presented here. Initially, there are six alleles observed in the pedigree. Braverman's method reduces the number of alleles to four, while the method presented here reduces the number of alleles to three. This example is originally from (Braverman, 1985).

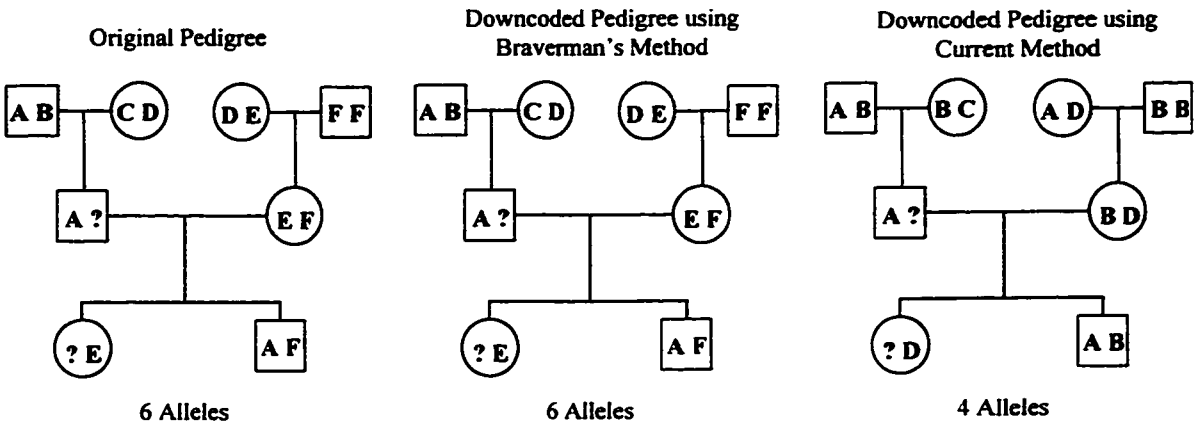


Figure 4.5 Example of the reduction in the number of alleles from downcoding using two methods.

This is an example that includes missing data in the pedigree. There were six alleles observed in the original pedigree. There was no reduction in the number of alleles using Braverman's method, while the method presented here reduced the number of alleles to four. This example is originally from (Braverman, 1985).

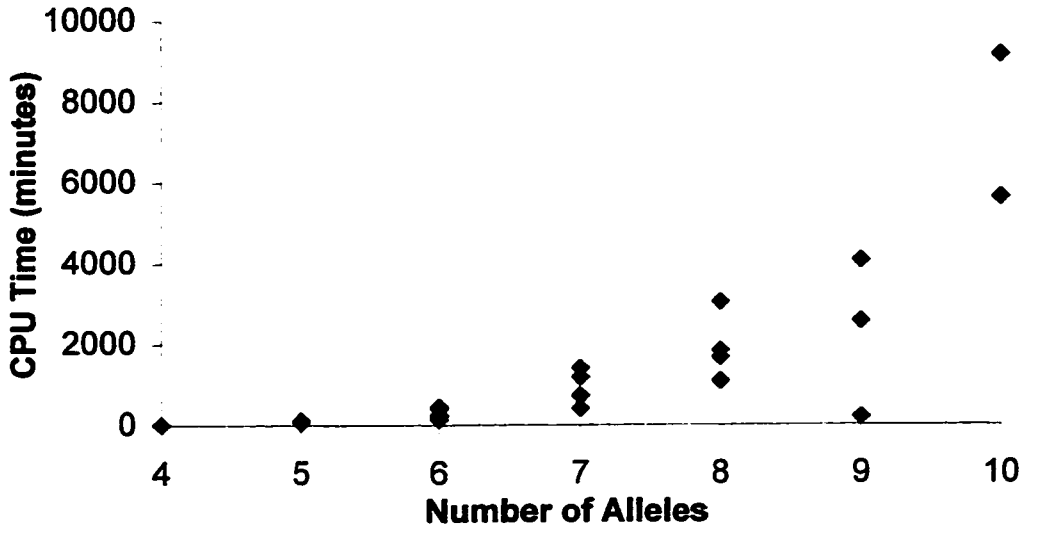


Figure 4.6 Effect of the number of marker alleles on the computation time for two-point linkage analysis.

As the number of alleles increases, the computation time also increases. Each point represents the computation time for a particular marker. Not all of the data points are independent since some markers were downcoded to determine the computation time with fewer alleles (see table 4.2). The pedigree that was used for these computations is shown in figure 4.7. All computations were performed using the program LIPED (Ott, 1974) on a Digital Alpha Station 200.

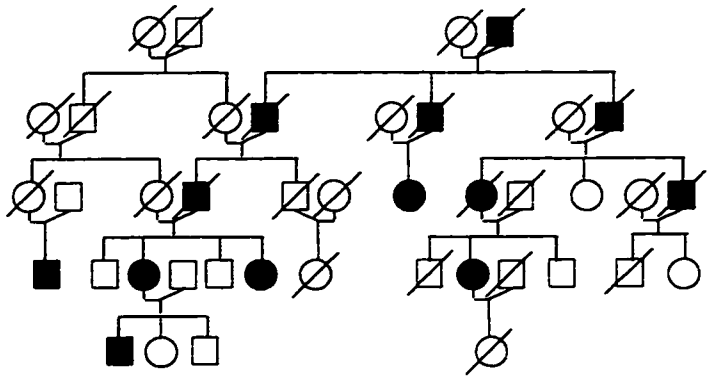


Figure 4.7 Pedigree structure used for the calculations in figure 4.6.

This pedigree was originally described in Bird et al., 1989, but the structure has been modified to include more individuals as shown above. A slash (/) through the symbol indicates that the individual does not have any marker genotype information. A darkened symbol indicates that the individual is affected.

Table 4.1 Computation times for two-point linkage analysis for the original and downcoded data.

The pedigree that was used for these computations is shown in figure 4.7. All computations were performed using the LIPED program (Ott, 1974) on a Digital Alpha Station 200.

Original Data		Downcoded Data	
Number of alleles	CPU time (minutes)	Number of alleles	CPU time (minutes)
10	9167	8	1672
10	5634	8	1089
8	3046	6	407
8	1846	7	729
8	1090	7	417
7	1410	5	122
6	254	5	76
6	227	5	66

Table 4.2 Computation times for multipoint linkage analyses using the original and downcoded data.

TIME: computation time using a Digital Alpha Station 200. LOD: lod scores computed at a recombination fraction of 0. NA= not possible to perform computations due to memory constraints. h = hours, m = minutes, s = seconds.

		Number of	LINKMAP		FASTLINK		VITESSE	
Example		Alleles/Locus	LOD	TIME	LOD	TIME	LOD	TIME
1	Original	5 5 5 2	2.08	67m. 16s.	2.08	6m. 22s.	2.08	< 5s.
	Downcoded	3 3 3 2	2.08	44s.	2.08	10s.	2.08	< 5s.
2	Original	5 5 4 3 2	1.308	48h. 17m.	NA	NA	1.307	< 5s.
	Downcoded	4 3 3 3 2	1.308	2h. 29m.	NA	NA	1.307	< 5s.

CHAPTER 5 : SUMMARY

We have evaluated several study design issues in order to identify cost-effective strategies for the gene mapping of complex traits. We evaluated issues relating to pedigree ascertainment schemes, marker maps and genotyping costs, and the computational efficiency of statistical methods. Evaluation of study design issues in terms of the overall cost of the design has been an effective strategy since we have been able to simultaneously consider multiple cost factors. Here, we describe our conclusions, and the main implications of our results.

First, we have demonstrated that different family configurations differ in the amount of information they provide to detect linkage. For both nuclear pedigrees and more extended pedigrees, we show that different sampling schemes that depend on the family configuration affect the resulting sample in terms of the fraction of individuals segregating for genes at different trait loci. This has an impact on the relative power to detect linkage among different selection schemes, which consequently affects the required sample size and the cost of the analysis. The ability to replicate linkage results may be affected by the selection schemes used among different studies. When linkage is detected in one study, failure to detect linkage in alternative samples may be a result of a difference in power to detect the original locus due to the ascertainment scheme despite a similar or larger sample size. Therefore, caution must be used in comparing results from studies that use different ascertainment schemes.

We were able to identify selection schemes that lower the cost of a study for a wide variety of underlying models. In particular, selection schemes that require at least three affected individuals are generally more cost-effective than selection schemes that require at least two affected individuals. The exceptions are 1) studies where only one ASP per sibship is used and the sporadic probability is low, and 2) studies to identify the lower dis-

ease prevalence locus in a two-locus model. Selection schemes that differ only in the number of unaffected individuals are usually not significantly different in terms of cost.

Second, we identified several characteristics of SNP markers that are required for a cost-effective genome screen compared to the current method of using STRP markers. We derived a statistic called MPIC for calculating the information content of a cluster of diallelic markers that is analogous to the single locus PIC. Using MPIC, we evaluated marker characteristics for clustered SNPs to identify a desirable range of model parameters including the allele frequency, the number of markers per cluster, the linkage disequilibrium, and the probability of being able to determine phase information. Not surprisingly, we found that all of the SNP map structures require more markers than the STRP map structure, and that SNPs can cost at most 60% of the cost per genotype for STRPs to be cost-effective. In the ideal case where the marker distances are known, we conclude that clustered SNP map structures are less efficient than a uniform SNP map structure based on the number of markers required for each map structure. If we do consider the effect of misspecifying the inter-marker distance, we found that using more informative markers such as the STRP markers or clusters of SNP markers will minimize the expected bias in the lod score when the markers are linked to the disease locus. Therefore, a better strategy than use of uniformly spaced SNP maps might be to use a genetic map with clusters of two SNP markers since this will reduce the bias in the lod score while only slightly increasing the cost compared to uniformly spaced SNP markers.

Finally, we demonstrated that the computational burden associated with linkage analysis can be reduced through the use of downcoding. For complex traits that often require large sample sizes, the use of computationally demanding analysis methods that can incorporate multiple markers and trait loci can have a significant impact on the overall cost of a study by reducing the amount of other resources such as pedigree collection or marker genotyping that are required to detect linkage. We presented a new method of downcoding that can be used for pedigrees with missing marker genotypes and with complex structure. This method is more efficient in reducing the number of alleles than previous

methods because it allows downcoding in portions of the pedigree that contain missing information, and it allows alleles within a mating to be relabeled to the same allele label. In addition, this method of downcoding can be used for a pedigree with arbitrary size and structure. Many of the other methods that are currently available for reducing the computation time for likelihood-based linkage analysis are restricted in the number of individuals, or the structure of the pedigree that can be evaluated. Downcoding broadens the potential analyses that are available for use with pedigree structures that do not fit the criteria of other methods.

The results presented in this dissertation suggest several additional questions. First, in our analysis of pedigree selection criteria, we did not compare the cost of ascertainment of extended pedigrees to nuclear pedigrees. Nuclear pedigrees are potentially easier and less expensive to collect per individual than extended pedigrees. In addition, we have demonstrated that for some models, pedigrees with fewer affected individuals are more informative for linkage within the context of a particular pedigree structure, suggesting that in some cases nuclear pedigrees may be more cost-effective than extended pedigrees. Comparing nuclear pedigrees with extended pedigrees is difficult because of the number of extended pedigree structures that can be considered. Also, the distribution of family size and structure is dependent on the population, which determines the relative probability of different pedigree structures. Finally, it may be difficult to extend the analytical method we used for nuclear pedigrees to larger pedigrees because of the many different relationships that are observed in extended pedigrees. A second extension of these results is to consider the use of SNP markers compared to STRP markers when methods other than multipoint linkage analysis are used. For example, the trade-off between the number of markers and the marker information may result in a different relative cost for the two marker types when association studies are considered.

BIBLIOGRAPHY

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

Amos CI, Krushkal J, Thiel TJ, Young A, Zhu DK, Boerwinkle E, de Andrade M (1997) Comparison of model-free linkage mapping strategies for the study of a complex trait. Genet Epidemiol 14:743-748

Bird TD, Sumi SM, Nemens EJ, Nochlin D, Schellenberg G, Lampe TH, Sadovnik A, et al (1989) Phenotypic heterogeneity in familial Alzheimer's disease: a study of 24 kindreds. Annals of Neurology 25(1):12-25

Bishop DT, Williamson JA (1990) The power of identity-by-state methods for Linkage analysis. Am J Hum Genet 46:254-265

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85-97

Bodmer WF, Bailey CJ, Bodmer J, Bussey HJ, Ellis A, Gorman P, Lucibello FC et al (1987) Localization of the gene for familial adenomatous polyposis on chromosome 5. Nature 328:614-616

Bonaventure J, Rousseau F, Legeai-Mallet L, Le Merrer M, Munnich A, Maroteaux P (1996) Common mutations in the fibroblast growth factor receptor 3 (FGFR3) gene account for Achondroplasia, Hypochondroplasia, and Thanatophoric Dwarfism. Am J Med Genet 63:148-154

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

Braverman MS (1985) An algorithm to improve the computational efficiency of genetic linkage analysis. Computers and Biomedical Research, 18:24-36

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and set-specific variation in recombination. Am J Hum Genet 63: 861-869

Brown PO, Hartwell L (1998) Genomics and human disease – variations on variation, Nature Genetics 18:91-93

Carey G, Williamson J (1991) Linkage analysis of quantitative traits: increased power by using selected samples. Am J Hum Genet 49:786-798

Carter BS, Beaty TH, Steinberg GD, Childs B, Walsh PC (1992) Mendelian inheritance of familial prostate cancer. Proc Natl Acad Sci 89:3367-3371

Chapman NH, Wijsman EM (1998) Sample size requirements for the detection of loci involved in complex disease: sib-pair methods vs. linkage disequilibrium testing, Am J Hum Genet 63(4):A227

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J et al (1996) Accessing genetic information with high-density DNA arrays, Science 274:610-614

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

Copeman JB, Cucca F, Hearne CM, Cornall RJ, Reed PW, Ronningen KS, Undlien DE et al (1996) Linkage disequilibrium mapping of a type 1 diabetes susceptibility gene (IDDM7) to chromosome 2q31-33. Nat Genet 9(1):80-85

Daniels JK, Spurlock G, Williams NM, Cardno AG, Jones LA, Murphy KC, Asherton P, et al (1997) Linkage study of chromosome 6p in sib-pairs with schizophrenia. Am J Med Genet 74(3):319-23

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, et al (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130-136

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

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, et al (1987) A genetic linkage map of the human genome. Cell 51(2):319-337

Durner M, Greenberg DA, Hodge SE (1992) Inter- and intrafamilial heterogeneity: effective sampling strategies and comparison of analysis methods. Am J Hum Genet 51:859-870

Eeles RA, Durocher F, Edwards S, Teare D, Badzioch M, Hamoudi R, Gill S, et al (1998) Linkage analysis of chromosome 18 markers in 136 prostate cancer families. Am J Hum Genet 62:653-658

Elston RC, Guo X, Williams LV (1996) Two-stage global search designs for linkage analysis using pairs of affected relatives. Genet Epidemiol 13:535-558

Field LL, Tobias R, Magnus T (1994) A locus on chromosome 15q26 (IDDM3) produces susceptibility to insulin-dependent diabetes mellitus. Nature Genet 8:189-194

Garner C, Kelley M, Cardon L, Joslyn G, Carey A, LeDuc C, Lichter J, et al (1996) Linkage analyses of schizophrenia to chromosome 6p24-p22: an attempt to replicate. Am J Med Genet 67(6):595-610

Ginsburg Ekh, Axenovich TI (1996) On planning of samples for linkage analysis: two ways of a sample size reduction. Genet Epidemiol 13:343-354

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, et al (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349:704-706

Goldgar DE, Easton DF (1997) Optimal strategies for mapping complex diseases in the presence of multiple loci. Am J Hum Genet 60:1222-1232

Goldin LR, Gershon ES (1988) Power of the affected sib-pair method for heterogeneous disorders. Genet Epidemiol 5:35-42

Goldin LR, Weeks DE (1993) Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. Am J Hum Genet 53:908-915

Greenberg DA, Hodge SE (1989) Linkage analysis under "random" and "genetic" reduced penetrance. Genet Epidemiol 6:259-264

Gu C, Todorov A, and Rao DC (1996) Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. Genet Epidemiol 13:513-533

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234-238

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684-1689

Hanis CL, Boerwinkle E, Chakroborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, et al (1996) A genome-wide search for human non-insulin-dependent

(type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nature Genetics 13:161-166

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behavior Genet 2(1):3-19

Hashimoto L, Habita C, Beressi JP, Delepine M, Besse C, Cambon-Thomsen A, Deschamps I, et al (1994) Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. Nature 371:161-164

Hasstedt S, Cartwright P (1981) PAP: Pedigree Analysis Package. Technical Report 13. Department of Medical Biophysics and Computing. University of Utah, Salt Lake City

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

Hodge SE (1984) The information contained in multiple sibling pairs. Genet Epidemiol, 1:109-122

The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72:971-983

Imperatore G, Hanson RL, Pettitt DJ, Kobes S, Bennett PH, Knowler WC (1998) Sib-pair linkage analysis for susceptibility genes for microvascular complications among Pima Indians with type 2 diabetes. Diabetes 47(5):821-30

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the Cystic Fibrosis gene: genetic analysis. Science 245:1073-1080

Kerem B, Kerem E (1996) The molecular basis for disease variability in Cystic Fibrosis. Eur J Hum Genet 4:65-73

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58: 1347-1363

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nature Genetics 17: 21-24

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci 84:2363-2367

Leppert M, Dobbs M, Scambler P, O'Connell P, Nakamura Y, Stauffer D, Woodward S, et al (1987) The gene for familial polyposis coli maps to the long arm of chromosome 5. Science 238:1411-1413

Levy-Lahad E, Wijsman EM, Nemens E, Anderson L, Goddard KA, Weber JL, Bird TD, et al (1995) A familial Alzheimer's disease locus on chromosome 1. Science 269:970-977

Lindblom A, Tannergard P, Werelius B, Nordenskjold M (1993) Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. Nature Genetics 5:279-282

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by In Vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397-401

Livak KJ, Marmaro J, Todd JA (1995) Towards fully automated genome-wide polymorphism screening. Nature Genetics 9:341-342

Luo D-F, Bui MM, Muir A, Maclaren NK, Thomson G, She JX (1995) Affected-sib-pair mapping of a novel susceptibility gene to insulin-dependent diabetes mellitus (IDDM8) on chromosome 6q25-q27. Am J Hum Genet 57:911-919

McCarthy MI, Kruglyak L, Lander ES (1998) Sib-pair collection strategies for complex diseases. Genet Epidemiol 15:317-340

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, et al (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266:66-71

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7(3):277-318

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616-1622

Nickerson DA, Kaiser R, Lappin S, Stewart J, Hood L, Landergren U (1990) Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. Proc Natl Acad Sci 87:8923-8927

Nickerson DA, Whitehurst C, Boysen C, Charmley P, Kaiser R, Hood L (1992) Identification of clusters of biallelic polymorphic sequence-tagged sites (STSs) that generate highly informative and automatable markers for genetic linkage mapping. Genomics 12:377-387

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nature Genetics 11:402-408

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

Ott J (1978) A simple scheme for the analysis of HLA linkages in pedigrees. Ann Hum Genet, 42:255-257

Ott J (1991) Analysis of Human Genetic Linkage, revised ed. The Johns Hopkins University Press, Baltimore

Parker SL, Tong T, Bolden S, Wingo PA (1996) Cancer Statistics, CA Cancer J Clin, 65:5-27

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Nat Acad Sci 91:5022-5026

Peltomäki P, Aaltonen LA, Sistonen P, Pylkkanin L, Mecklin JP, Jarvinen H, Green JS, et al (1993) Genetic mapping of a locus predisposing to human colorectal cancer. Science 260:810-812

Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6:133-138

Pulver AE, Lasseter VK, Kasch L, Wolyniec P, Nestadt G, Blouin JL, Kimberland M, et al (1995) Schizophrenia: a genome scan targets chromosomes 3p and 8p as potential sites of susceptibility genes. Am J Med Genet 60(3):252-60

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, et al (1989) Identification of the Cystic Fibrosis gene: cloning and characterization of the complementary DNA. Science 245:1066-1073

Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229-241

Risch N, Giuffra L (1992) Model misspecification and multipoint linkage analysis. Hum Hered 42:77-92

Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584-1589

Risch N, Zhang H (1996) Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. Am J Hum Genet 58:836-843

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

Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, et al (1989) Identification of the Cystic Fibrosis gene: chromosome walking and jumping. Science 245: 1059-1065

Saiki RK, Walsh PS, Levenson CH, Erlich HA (1989) Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. Proc Natl Acad Sci 86:6230-6234

Schellenberg GD, Bird TD, Wijsman EM, Orr HT, Anderson L, Nemens E, White JA, et al (1992) Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. Science 258:668-671

Schellenberg GD, Payami H, Wijsman EM, Orr HT, Goddard KA, Anderson L, Nemens E, et al (1993) Chromosome 14 and late-onset familial Alzheimer disease (FAD). Am J Hum Genet 53:619-628

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127-1136

Sham PC, Zhao JH, Curtis D (1997) Optimal weighting scheme for affected sib-pair analysis of sibship data. Ann Hum Genet 61:61-69

Sribney WM, Swift M (1992) Power of sib-pair and sib-trio linkage analysis with assortative mating and multiple disease loci. Am J Hum Genet 51:773-784

Stine OC, McMahon FJ, Chen L, Xu J, Meyers DA, MacKinnon DF, Simpson S, et al (1997) Initial genome screen for bipolar disorder in the NIMH genetics initiative pedigrees: chromosomes 2, 11, 13, 14, and X. Am J Med Genet 74(3):263-9

Suarez BK, Rice J, Reich T (1978) The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87-94

Suarez BK, Van Eerdewegh P (1984) A comparison of three affected-sib-pair scoring methods to detect HAL-linked disease susceptibility genes. Am J Med Genet 18:135-146

Suarez BK, Hampe CL, Van Eerdewegh P (1994) Problems of replicating linkage claims in psychiatry. In: Gershon, Cloninger (eds) Genetic approaches to mental disorders. American Psychiatric Press, Washington DC, pp 23-46

Terwilliger JD, Ding Y, Ott J (1992) On the relative importance of marker heterozygosity and intermarker distance in gene mapping. Genomics 13:951-956

Tsui L-C, Buchwald M, Barker D, Braman JC, Knowlton R, Schumm JW, Eiberg H, et al (1985) Cystic Fibrosis locus defined by a genetically linked polymorphic DNA marker. Science 230:1054-1057

Vahava O, Morell R, Lynch ED, Weiss S, Kagan ME, Ahituv N, Morrow JE, et al (1998) Mutation in transcription factor POU4F3 associated with inherited progressive hearing loss in humans. *Science* 279:1950-1954

Van Duijn CM, Hendriks L, Farrer LA, Backhovens H, Cruts M, Wehnert A, Hofman A, et al (1994) A population-based study of familial Alzheimer disease: linkage to chromosomes 14, 19, and 21. Am J Hum Genet 55:714-727

Vieland V, Greenberg DA, Hodge SE, Ott J (1992a) Linkage analysis of two-locus diseases under single-locus and two-locus analysis models. Cytogenet Cell Genet 59:145-146

Vieland VJ, Hodge SE, Greenberg DA (1992b) Adequacy of single-locus approximations for linkage analyses of oligogenic traits. Genet Epidemiol 9:45-59

Vieland VJ, Greenberg DA, Hodge SE (1993) Adequacy of single-locus approximations for linkage analysis of oligogenic traits: extension to multigenerational pedigree structures. Hum Hered 43:329-336

Vogler GP, Wette R, McGue MK, Rao DC (1995) Properties of alternative estimators of familial correlations under variable sibship size. Biometrics 51:276-283

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077-1082

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388-396

Wijsman EM (1997) Association vs. linkage analysis in mental disorders. In: Blum K, Noble ER (eds) Handbook of Psychiatric Genetics. CRC Press, Boca Raton, FL

Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, et al (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 265:2088-2090

Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, et al (1995) Identification of the breast cancer susceptibility gene BRCA2. Nature 378:789-792

Wu DY, Ugozzoli L, Pal BK, Wallace RB (1989) Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. Proc Natl Acad Sci 86:2757-2760

Zamani M, Pociot F, Raeymaekers P, Nerup J, Cassiman JJ (1996) Linkage of type I diabetes to 15q26 (IDDM3) in the Danish population. Hum Genet 98:491-496

Ziegle JS, Su Y, Corcoran KP, Nie L, Mayrand PE, Hoff LB, McBride LJ, et al (1992) Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics 14:1026-1031

VITA

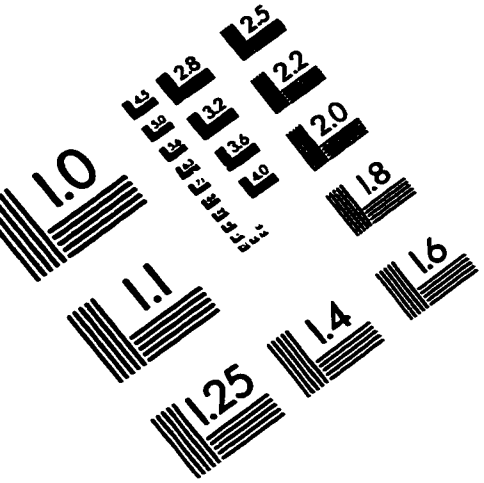
Katrina Blouke Goddard

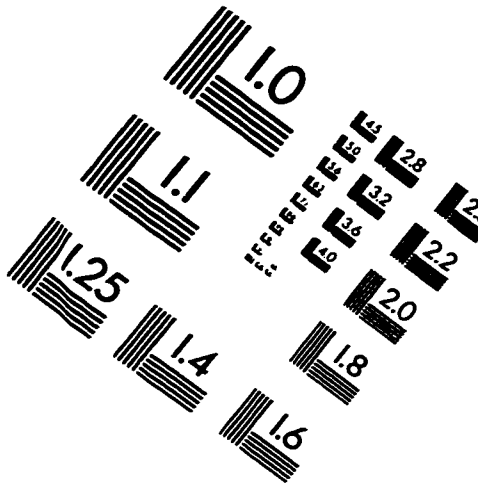
University of Washington

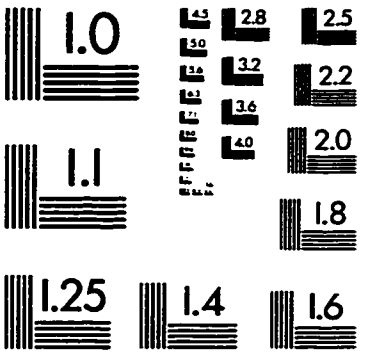
1999

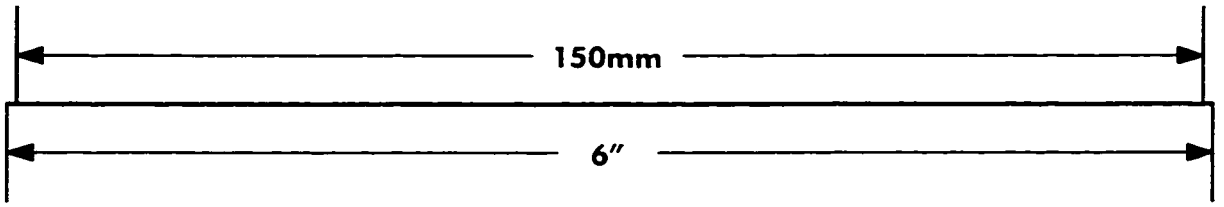
Katrina Blouke Goddard received a Bachelor of Science degree with Honors in Molecular Biology from the University of Wisconsin-Madison in 1990. She received a Masters of Science in Biostatistics from the University of Washington in 1995. Since then, she has been enrolled in the Biostatistics Ph.D. program at the University of Washington.

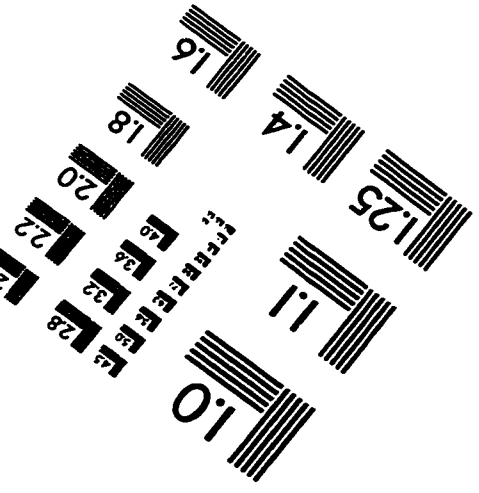
IMAGE EVALUATION
TEST TARGET (QA-3)

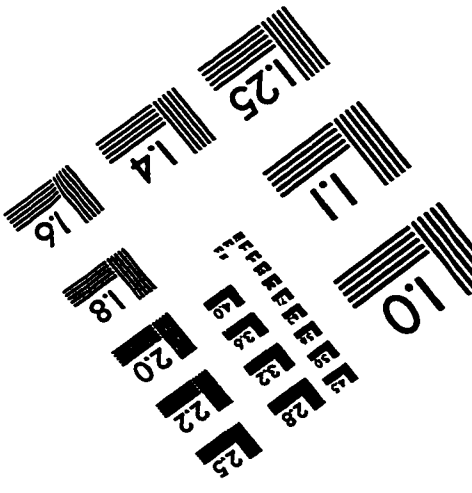












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989