EFFICIENT CONFIDENCE SETS FOR DISEASE GENE LOCATIONS

by

RITWIK SINHA

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Epidemiology and Biostatistics
CASE WESTERN RESERVE UNIVERSITY

May, 2007
We hereby approve the dissertation of

Ritwik Sinha

candidate for the Ph. D. Degree*.

(Signed)

Yuqun Luo, Ph. D
Department of Epidemiology and Biostatistics
Chair of the committee

Robert C. Elston, Ph. D
Department of Epidemiology and Biostatistics

Katrina A. B. Goddard, Ph. D
Department of Epidemiology and Biostatistics

J. Sunil Rao, Ph. D
Department of Epidemiology and Biostatistics

John Barnard, Ph. D
Cleveland Clinic Foundation

(date) February 7, 2007

*We also certify that written approval has been obtained for any proprietary material therein.
To Moumita, Ma and Baba
Contents

1 Introduction 3

1 Literature Review: Current Approaches To Confidence Set Construction 9

2 Model Free Linkage Analysis 10

  2.1 Proportion Test and Mean test 12

  2.2 Likelihood Based Model Free Linkage Analysis 13

     2.2.1 Maximum LOD Score Criterion for Affected Relative Pairs 14

  2.3 MLS Under Possible Triangle Constraints 18

  2.4 A General Conditional Logistic Model for Affected Relative Pair Linkage Analysis 20

  2.5 Linkage Tests for Affected Pedigree Members 22

  2.6 Comparison of Model Free Linkage Statistics 23

3 Confidence Set Inference 26
4.2.1 The Approach of Liang et al. [2001a] .......................... 52
4.2.2 Extensions ...................................................... 54
4.2.3 Coverage of Confidence Intervals ............................ 54
4.3 Comparison of Intermediate Mapping Methods ............... 55
  4.3.1 Methods ....................................................... 56
  4.3.2 Simulation and Results ...................................... 57

5 Bayesian Methods in Linkage ................................. 60
  5.1 Principles of Bayesian Statistics ............................... 61
  5.2 Bayesian Methods in Linkage Analysis ........................ 62
  5.3 Bayesian Credible Intervals for Disease Gene Locations .... 64
    5.3.1 Linkage Disequilibrium Based Fine Mapping .......... 64
    5.3.2 Quantitative Trait Locus Mapping .................... 65

II Two Novel Approaches To Construct Efficient Confidence Sets 66

6 Confidence Set Inference with Likelihood Ratio Test ........ 67
  6.1 More Efficient CSI with Likelihood Ratio Tests: CSI-MLS .... 68
    6.1.1 Single-point CSI-MLS .................................. 69
    6.1.2 Multi-point CSI-MLS .................................... 71
    6.1.3 Powers of CSI-MLS and CSI-Mean .................... 73
8.3 Discussion .......................................................... 115

9 Application of Proposed Methods to Real Data 126

9.1 Data and Methods ............................................... 127

9.2 Results .............................................................. 128

Appendices ............................................................ 130

A Confidence Set Inference for the Mean Test 130

A.1 Stochastic Ordering of Mean Test Statistic ................... 130

B Proofs ................................................................. 135
List of Tables

3.1 Two-IBD test statistic properties ................................. 36

6.1 Single locus disease models ........................................... 85

6.2 Two-locus disease models ............................................. 86

6.3 Parameters of two-locus disease models ......................... 87

6.4 Multi-point CSI on MS markers ................................. 88

6.5 Multi-point CSI on SNP markers ................................. 89

6.6 Multi-point CSI for two-locus disease models ................. 90

6.7 Power of CSI-MLS and CSI-Mean ................................. 91

7.1 Mean Squared Errors of the estimates step one ................. 99

7.2 Mean lengths of CSI-MLS and CSI-Mean ......................... 100

8.1 Bayesian sensitivity analysis ........................................... 120

8.2 Single locus disease models ........................................... 120

8.3 Length and coverage of intermediate fine mapping methods .... 122

8.4 Length and coverage for two-locus models ....................... 123
8.5 Length and coverage of intermediate fine mapping methods with missing parental genotypes ........................................ 124
8.6 RMSE of location estimates ........................................ 124
8.7 Empirical coverage of confidence intervals with variation of Bayesian interval 125
8.8 Drop size to have 95% empirical coverage ................... 125
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Triangle constraints</td>
<td>19</td>
</tr>
<tr>
<td>4.1</td>
<td>LOD support interval</td>
<td>50</td>
</tr>
<tr>
<td>6.1</td>
<td>Single point CSI</td>
<td>84</td>
</tr>
<tr>
<td>6.2</td>
<td>Empirical power curves</td>
<td>85</td>
</tr>
<tr>
<td>6.3</td>
<td>Power CSI-Mean and CSI-MLS over range of disease models</td>
<td>86</td>
</tr>
<tr>
<td>6.4</td>
<td>Contour plots of Power difference between CSI-Mean and CSI-MLS</td>
<td>87</td>
</tr>
<tr>
<td>7.1</td>
<td>Estimates of $z_1$ and $z_2$</td>
<td>99</td>
</tr>
<tr>
<td>8.1</td>
<td>Probability density function for chosen Beta distribution</td>
<td>117</td>
</tr>
<tr>
<td>8.2</td>
<td>Convergence of MCMC</td>
<td>118</td>
</tr>
<tr>
<td>8.3</td>
<td>Intermediate fine mapping methods for single-locus disease model</td>
<td>119</td>
</tr>
<tr>
<td>8.4</td>
<td>Intermediate fine mapping methods for two-locus disease model</td>
<td>121</td>
</tr>
<tr>
<td>9.1</td>
<td>Application of proposed methods to real data</td>
<td>129</td>
</tr>
</tbody>
</table>
I thank my advisor, Dr. Yuqun Luo, for her time and guidance in making this dissertation a reality. I also thank Dr. Robert C. Elston, Dr. Katrina A. B. Goddard, and Dr. J. Sunil Rao for serving on the committee that evaluated my work, and for providing me with the opportunity to bother them with my questions. I am grateful to Dr. John Barnard for agreeing to serve on my committee at a short notice. Many discussions with Dr. Sudha Iyengar, Dr. Shili Lin, and Dr. Catherine Loader have helped improve this dissertation. Sincere thanks go to them and to my friend Shiv Saini for helping me with parts of this dissertation. Numerous teachers, who have imparted knowledge to me through the years, have helped shape this dissertation. I would like to record my gratitude to all of them.

I am grateful to my friends at Case Western Reserve University for helping me better understand Genetic Epidemiology and making my time here enjoyable. In particular, the names Qing Lu, Murielle Bochud, Moumita Sinha, Douglas Londono, Tao Wang, Chao Xing, Guan Xing, Sungho Won, Mohammed Orloff, Yeunjoo Song, and Gyungah Jun come to mind.

This dissertation would not have been possible without the people who release their software with open source. In particular, I would like to thank R, Linux, Allegro, Merlin, and CSI. The software suite S.A.G.E has been used in parts of this dissertation.

Finally, I thank my family for their love and support.
Efficient Confidence Sets for Disease Gene Locations

Abstract

by

RITWIK SINHA

In positional cloning of disease susceptibility genes, identification of a linked chromosomal region via linkage studies is often followed by fine mapping with association studies. Efficiency can be gained via an intermediate step where confidence regions for the locations of disease genes are constructed. We proposed and explored the properties of two novel practical approaches, one frequentist and one Bayesian, to constructing such intervals using affected sibling pair data.

The first approach draws upon a promising paradigm, Confidence Set Inference (CSI) [Papachristou and Lin, 2006a], that converts a sequence of tests to obtain the interval. CSI replaces the traditional null hypotheses of no linkage with a new set of null hypotheses where the chromosomal position under consideration is in tight linkage with a trait locus, and was proposed for the Mean test statistics (CSI-Mean). We postulate that a more efficient test statistic, the Maximum LOD Score (MLS), will lead to more efficient confidence sets when used in the CSI framework. We propose a procedure that tests the CSI null hypotheses using the MLS statistic (CSI-MLS). Compared to CSI-Mean, CSI-MLS provides tighter confidence regions over a range of single- and two-locus disease models. In addition, the MLS test is more powerful than the Mean test in testing the CSI null over a wide range of disease models. Furthermore, CSI-MLS is computationally much more efficient than CSI-Mean. The CSI framework requires knowledge of some disease model related parameters. In practice, such knowledge is often absent and a two-step procedure may be employed. The advantages of CSI-MLS over CSI-Mean is preserved in this practical setting as well.
Though the CSI framework compares favorably with other competitors, and CSI-MLS has further improved its statistical and computational efficiency, the two-step procedure is often conservative. This motivates us to explore a Bayesian approach, that formulates the disease gene location as a parameter, to seek possible improvements. In this case, credible intervals with a uniform prior on the location are confidence regions. A Metropolis-Hastings algorithm is implemented to sample from the posterior distribution and Highest Posterior Density Intervals of the disease gene location are constructed. The proposed Bayesian method is shown to provide precise confidence sets with correct coverage probabilities when compared to competing methods. The two novel methods are applied to a Rheumatoid Arthritis data example.
Chapter 1

Introduction

Before the 1980s, *Functional Cloning* was the predominant paradigm in dissecting the genetic causes of diseases, where proteins with known functions that might be involved in the pathway leading to the disease is the point of entry in the inquiry. Over the last three decades, a new paradigm, *Positional Cloning*, has proved very successful in the quest for the underlying susceptibility genes in Mendelian disorders. This new paradigm, made possible by the tremendous advances in genotyping technology and computational capacity, starts with narrowing down the promising genomic regions using the abundant genetic markers along the human chromosomes as signposts [Collins, 1995]. This is then followed up by searching databases for candidate genes within the identified region, a step that has become ever more promising since the completion of a draft sequence of the human genome [Lander et al., 2001]. Finally, molecular work will be applied to the candidate genes, or if none is available, sequencing will be employed to identify the culprits.

In order for the cost of the molecular work to be affordable, it is critical to identify as narrow a genomic region as possible in the initial steps, yet not too narrow to miss the genes. The initial identification of the genomic region is often through linkage studies on family data. Two genetic loci tend to co-segregate more often when they are closer to each other, and
this distance is measured by the *Recombination fraction* between the two, the probability that the two alleles at a locus of an individual are inherited from different grandparental copies of one of the parents. While this is a cost effective way of scanning the whole human genome for disease genes, owing to the limit on sample sizes that are practical the resolution of a linkage study is often around 10 centi Morgans, approximately 10 Million base pairs [Reeve and Ranala, 2002], within which around 300 genes may reside. Thus, a fine mapping strategy often follows the linkage identification effort. Fine mapping is based on the population phenomenon of linkage disequilibrium (LD), where alleles at two genetic loci very close together show statistical association at the population level. Ideally, LD arises when a disease mutation occurred many generations ago on a chromosome and thus was exclusively associated with the marker alleles residing on that same chromosome. As time goes by, recombination events will break down this association, the rate of decay being geometrically related to the genetic distance. Since many more recombination events have occurred compared to what can be hoped for within the families used for linkage study, LD will persist only between loci that are much closer. Thus LD based fine mapping can provide a resolution in the order of 0.5 cM or less.

In the past, linkage studies had focused on hypothesis testing to identify regions that can be claimed to be linked to the trait under consideration. In the absence of an interval estimate of the susceptibility gene location, ensuing fine mapping efforts may not reap the maximum benefit from information that is contained in the data traditionally used for linkage study. In the past few years, researchers have recognized that the efficiency in narrowing down genomic regions via a genome-wide linkage scan followed by LD mapping can be improved by inserting an intermediate step that provides an interval estimator with sound statistical properties [Papachristou and Lin, 2006c]. The goal of this intermediate step is to localize disease genes to narrow regions of the genome with controlled confidence using dense marker maps on families. In addition to making the follow up association studies more efficient and cost effective, it helps better address the multiple testing issues.
Several lines of investigations have since been pursued, which we shall review briefly here, and in more detail in the section on literature review.

While the problem of supplying a confidence interval for a Quantitative Trait Locus (QTL) has been extensively studied in experimental organisms, where crosses can be controlled (e.g., Lander and Botstein [1989], Mangin et al. [1994], Visscher et al. [1996], Dupuis and Siegmund [1999]), its investigation in humans poses additional challenges (e.g., Liang et al. [2000]). For binary traits in humans, LOD-support intervals are often employed to construct confidence intervals of disease gene locations using data from linkage studies [Kristjansson et al., 2002]. However, the statistical properties of such intervals are not clear [Nemesure et al., 1995, Dupuis and Siegmund, 1999, Papachristou and Lin, 2006c]. Focusing on ASPs, Liang et al. [2001a] proposed a Generalized Estimating Equations (GEE) approach to simultaneously estimate susceptibility gene locations and their effect sizes.

Parallelly, the Confidence Set Inference (CSI), based on the duality of confidence set construction and hypothesis testing, was developed [Lin et al., 2001, Lin, 2002]. CSI replaces the traditional null hypotheses of no linkage with a new set of null hypotheses where the chromosomal position under consideration is in tight linkage with a trait locus. The Mean test statistic was reformulated to test the CSI hypothesis (CSI-Mean) [Papachristou and Lin, 2005a, 2006a]. Contrary to the GEE approach’s simultaneous estimation of disease model related parameters and the disease gene location, the CSI approaches require knowledge of certain parameters related to the disease model, relative risks of siblings and offspring being one possibility. A two-step approach that precedes CSI with the estimation of these disease parameters was shown to perform well [Papachristou and Lin, 2006b,c]. In addition to the above three widely used or more fully developed approaches, there exists other less studied methods (e.g., Hauser et al. [1996], Hössjer [2003]).

Recent studies have shown that the confidence intervals produced by the GEE method might have reduced coverage [Lebrec et al., 2006, Papachristou and Lin, 2006c], the bias
being severe in regions where the allelic Identity by Descent (IBD) probability estimates are less precise. In a comparison of the GEE, CSI-Mean, LOD-support and two bootstrap methods, Papachristou and Lin [2006c] showed that when the observed coverage probability was held to be the same, the intervals obtained by the CSI were better able to localize the disease causing loci.

Encouraged by the sound statistical properties of the CSI based intermediate fine mapping approaches, we explore other test statistics to improve their efficiency and to reduce the computational burden. CSI-Mean is based on the Mean test, which compares the mean observed number of alleles shared IBD at a genomic location to what is expected under the CSI null, that this location is in tight linkage with a trait locus. In its current implementation, CSI-Mean requires Monte Carlo simulations to estimate the mean and the variance of the test statistic, and thus is computationally demanding. Since CSI is based on the idea of converting a group of hypothesis tests at various genomic locations to obtain the confidence sets for the true disease gene location(s), we postulate that more powerful test procedures should lead to more efficient estimators of confidence sets. The Maximum LOD Score (MLS) test [Risch, 1990c] is a likelihood ratio test (LRT) in traditional linkage analysis that compares the distribution of the IBD sharing to that expected under the null of no linkage, with incomplete marker information handled naturally. Linkage tests based on MLS have been shown to perform well over a range of disease models [Davis and Weeks, 1997]. Thus, we propose to reformulate the MLS statistic in the CSI framework (CSI-MLS). An immediate benefit is the elimination of the need of Monte Carlo simulation and thus the improvement in computational efficiency.

In this thesis we will first derive the CSI-MLS procedures, both single-point and multi-point. Extensive simulations under disease models of varying complexity are performed to compare the efficiency of CSI-Mean and CSI-MLS in terms of coverage probabilities and precision of the confidence sets. The CSI framework requires knowledge of some disease
model related parameters. These parameters can possibly be estimated from epidemiologic studies. First, we shall assume that the required parameters are known, and compare CSI-MLS and CSI-Mean on the effect of a number of interesting factors to gain a comprehensive understanding. In particular, the space of potential disease models over which CSI-MLS outperforms CSI-Mean is identified. In the absence of knowledge of the disease model related parameters, Papachristou and Lin [2006b] proposed a two-step procedure: the required disease parameters are estimated in the first step and used in the second step to construct the confidence set. We show next that the CSI-MLS as implemented in the two-step procedure still performs better than CSI-Mean.

Though the CSI framework compares favorably with other competitors, and CSI-MLS has further improved its statistical and computational efficiency, the two-step procedure is often conservative [Papachristou and Lin, 2006b]. This motivates us to explore a Bayesian approach, that formulates the disease gene location as a parameter, to seek possible improvements. In this case, credible intervals with a uniform prior on the location are confidence regions. A Metropolis-Hastings algorithm is implemented to sample from the posterior distribution and Highest Posterior Density Intervals of the disease gene location are constructed. The proposed Bayesian method is compared to the two-step CSI-MLS, the LOD-drop approach, the GEE approach, and a Bootstrap method, over a range of simulated single and two-locus disease models. The proposed Bayesian method is shown to provide precise confidence sets with correct coverage probabilities when compared to competing methods. The two novel methods, together with other competitors, are applied to a Rheumatoid Arthritis data example.

The dissertation is organized into two main parts: literature review (chapters 2 through 5); and the two novel approaches and their properties (chapters 6 through 9). Since CSI is a general framework whose hypotheses can be tested by reformulation of any reasonable traditional linkage test statistic, chapter 2 reviews some of these possibilities, including
the MLS method [Risch, 1990c]. Existing methods for constructing confidence sets for disease gene locations are reviewed in chapters 3 and 4, followed by a review of Bayesian statistics and its applications to genetic epidemiology in chapter 5. In Chapters 6 and 7 we develop the CSI-MLS approach and study its properties. The Bayesian alternative to constructing credible sets for disease gene locations follows. Chapter 9 presents an application of the two novel methods to a real data example. As an aside, the single-point CSI-Mean proposed by [Lin, 2002, Papachristou and Lin, 2005a] requires that the Mean test statistic be stochastically decreasing in its distance from the disease locus. While it is an intuitive property, we provide a rigorous proof of this property in Appendix A.
Part I

Literature Review: Current Approaches
To Confidence Set Construction
Chapter 2

Model Free Linkage Analysis

Model-free linkage methods, in contrast to model-based linkage methods, do not depend on prior specification of a mode of inheritance for the disease or trait of interest. In other words, the frequencies and penetrances of disease genotypes does not have to be known in advance. Generally, model-free linkage analysis methods are applied to traits believed to have a complex genetic etiology - multiple genetic contributors that may follow different modes of inheritance and may interact with one another and or with environmental factors. For such traits it is difficult, if not impossible, to determine an adequate mode of inheritance prior to performing a search for linked loci.

Model-free methods of linkage analysis were first derived for samples comprising sib pairs by Penrose [1935] and Haseman and Elston [1972]. In addition to the advantage that the mode of inheritance of the disease need not be known in advance, sib-pair methods are easy to use and robust to ascertainment. Also, sib pairs are often the most frequently encountered affected relatives. Indeed, for diseases of complex etiology and late onset, there are often no large, multiply affected pedigrees. Certain sib pair linkage analysis methods can be extended to include information from other types of relatives [Risch, 1990c, Olson, 1999] or use more general sets of families [Whittemore and Halpern, 1994].
Affected relative pair (ARP) linkage analysis for qualitative traits involves testing if the allele sharing at a marker locus is more than would be expected if the marker was segregating independently of the disease. Two alleles are said to be Identical by Descent (IBD) if they have descended from the same ancestral allele. For example, two siblings can share either 0, 1 or 2 alleles IBD. Most other relative types can share at most 1 allele IBD in the absence of inbreeding.

A large number of affected relative pair tests are based on the number of alleles shared IBD by relative pairs [Day and Simmons, 1976, Green and Woodrow, 1977, Blackwelder and Elston, 1985, Risch, 1990c, Holmans, 1993, Olson, 1999]. Risch [1990c] proposed a maximum likelihood method for detecting linkage to complex diseases using affected sibling pairs. It was one of the first Affected Sibling Pair (ASP) linkage methods proposed that could successfully tackle situations where the exact IBD sharing is not known. Also, being a likelihood based approach it can be interpreted as a LOD score.

Whittemore and Halpern [1994] proposed a class of affected relative pair linkage analyses. These tests use the inheritance vector, a vector that quantifies the allele sharing in families and contains more information than the IBD sharing statistic in pairs, IBD sharing being a function of the inheritance vector. The tests are given by scoring inheritance vectors, different scoring functions giving rise to different linkage statistics. All these tests assume that there is no distortion to Mendelian segregation ratios, which implies that randomly selected siblings will share 0, 1 and 2 alleles with probabilities 1/4, 1/2 and 1/4, respectively.

For quantitative traits, Haseman and Elston [1972] suggested a regression method. In the original Haseman-Elston regression, the squared trait difference is regressed on the number of alleles shared IBD at a marker. A significant negative slope suggests linkage of the trait to the marker. In this Chapter I will restrict my attention only to linkage methods specifically developed for binary traits, namely, the Proportion and Mean tests [Blackwelder and Elston, 1985], the MLS test [Risch, 1990c] and the NPL test [Whittemore and Halpern,
2.1 Proportion Test and Mean test

Day and Simmons [1976] and Green and Woodrow [1977] proposed two simple statistics to test for evidence of linkage. One tests if the number of affected sib pairs sharing 2 alleles IBD is greater than the null expectation of 1/4 [Day and Simmons, 1976]; this test is given the names “Proportion” test and “Two-IBD” test. The other tests if the mean of the number of alleles shared IBD is greater than the null mean of 1 [Green and Woodrow, 1977], this test is referred to as the “Mean” test. Blackwelder and Elston [1985] compared generalized versions of the Proportion and Mean test to a $\chi^2$ goodness of fit test. The significance level and power are compared for known values of the population prevalence of the disease $K$, penetrance vector and recombination fraction $\theta$.

Let $r_{jk}$ be the observed number of sib pairs with $k$ marker alleles IBD and $j$ affected members and let $n_j = r_{j0} + r_{j1} + r_{j2}$ be the number of sib pairs with $j$ affected members; it is assumed that the exact number of alleles shared IBD is known. The parameters of interest are $p_{jk}$, the probability that a sib pair with $j$ affected members shares $k$ marker alleles IBD. The probability $p_{jk}$ is estimated by $\hat{p}_{jk} = r_{jk}/n_j$.

The Proportion test tests the following hypothesis, $H_0 : p_{22} = p_{12}$ against $H_1 : p_{22} > p_{12}$ (if $n_1 \neq 0$) or $H_0 : p_{22} = 1/4$ against $p_{22} > 1/4$ (if $n_1 = 0$). The test statistics used are as follows, if $n_1 \neq 0$, $t_1 = (\hat{p}_{22} - \hat{p}_{12})/s_1$, $s_1^2 = (1/n_1 + 1/n_2)3/16$ and if $n_1 = 0$, $t_1 = (\hat{p}_{22} - 1/4)/s_1$, $s_1^2 = 3/16n_2$.

The Mean test has the following null and alternative hypothesis. $H_0 : p_{21} + 2p_{22} = p_{11} + 2p_{12}$ against $H_1 : p_{21} + 2p_{22} > p_{11} + 2p_{12}$ (if $n_1 \neq 0$) and $H_0 : p_{21} + 2p_{22} = 1$ against $H_1 : p_{21} + 2p_{22} > 1$ (if $n_1 = 0$). The corresponding test statistics are: if $n_1 \neq 0$,
\[ t_2 = (\hat{p}_{21} + 2\hat{p}_{22} - \hat{p}_{11} + 2\hat{p}_{12}) / s_2, \text{ where } s_2^2 = (1/n_1 + 1/n_2)/2, \text{ and if } n_1 \neq 0; \ t_2 = (\hat{p}_{21} + 2\hat{p}_{22} - 1) / s_2, \text{ where } s_2^2 = 1/2n_2. \]

For sufficiently large values of \( n_1 \) and \( n_2 \), \( t_1 \) and \( t_2 \) can be assumed to have standard normal distributions under the null hypothesis.

Often when an ASP is ascertained it is informative to include in the sample known unaffected sibs in the same sibship. This article argues, when there is no linkage, sibships of size three or greater can be treated as collections of independent sib pairs and can be included without altering the null distribution of the test statistics. Hodge [1984] showed that within each sibship, under the null hypothesis of no linkage, the number of marker alleles shared IBD are pairwise (not mutually) independent. This pairwise independence guarantees that \( t_1 \) and \( t_2 \) can be used treating sibships as collections of independent pairs.

It is shown that the Mean test has an acceptable level of significance and is generally more powerful than the other two statistics. Further, Knapp et al. [1994a] determined that, provided \( \delta_1^2 = \delta_0 \delta_2 \) (as is true for some recessive models) where \( \delta_0 = P(\text{affected|trait genotype with } s \text{ susceptibility alleles}) \), the mean test is uniformly most powerful in \( \theta \), the recombination fraction between trait and marker loci.

### 2.2 Likelihood Based Model Free Linkage Analysis

Haseman and Elston [1972] considered linkage analysis of quantitative traits with markers of incomplete polymorphism, that is, where the exact IBD sharing at the marker is not known. However, for binary traits, this problem was studied in some detail in the late 80’s and early 90’s. This problem is especially acute if parents are unavailable for typing, as is often the case for late onset diseases. One approach to handle markers with incomplete polymorphism is to use identity by state (IBS) based linkage methods [Lange, 1986, Bishop
and Williamson, 1990, Green and Greenan, 1991]. Another approach is the likelihood based method suggested by Risch [1990c]. The main difference between Risch [1990c] and Bishop and Williamson [1990] is that Risch’s method takes into account not only how many alleles an affected pair share but also which ones they are. This also means that a pair sharing rare alleles IBD will contribute more evidence for linkage than will a pair sharing a common allele. Risch’s method utilizes more marker sharing information and should give us a more powerful test. Some popular likelihood based affected relative pair linkage analysis methods are reviewed in the following sections.

### 2.2.1 Maximum LOD Score Criterion for Affected Relative Pairs

The Maximum LOD Score (MLS) [Risch, 1990c] method was one of the earliest likelihood based affected relative pair (ARP) linkage analysis methods. The series of papers by Risch [1990a,b,c] is an important contribution to the study of ARP linkage analysis. In addition to proposing the MLS statistic, these papers provide a number of quantities important to the study of ARP linkage studies.

Risch [1990a] studies the parameters $\lambda_R$, the relative risk of disease to a relative of type R of an affected individual. The mathematical relationship between $\lambda_R$ for various relationships is computed for single and multilocus models.

Define a random variable $X_1$ to be 1 if individual 1 is affected and 0 otherwise. Similarly, define $X_2$ for individual 2, a relative of individual 1 of type R. Then the population prevalence is given by, $K = E(X_1)$ and the recurrence risk for a type R relative of an affected individual is $K_R = E(X_2|X_1 = 1)$. Define $\lambda_R$ as the risk ratio for a type R relative of an affected individual compared with population prevalence,

$$\lambda_R = \frac{K_R}{K}.$$
The parameters $\lambda_R$, for various relationship types, are provided by Risch [1990a] in terms of the additive genetic variance ($V_A$) and the dominance genetic variance ($V_D$) [James, 1971].

Risch [1990b] explored the IBD sharing probability at locations linked to a disease gene and studied the type of relative pairs that provide the most power for linkage studies.

Let the null IBD sharing probabilities for a random relative pair of type R, at a marker be given by $\alpha_{Ri} = P(\text{IBD} = i \text{ for relative of type R})$, where $\text{IBD}$ is the number of alleles shared identical by descent by the pair under consideration at the loci. First, consider the case when recombination between marker and trait locus is zero. Then the IBD sharing probabilities at the disease locus, $z_{Ri} = P(\text{IBD} = i | \text{affected relative pair of type R})$ are given by the equation

$$z_{Ri} = \frac{P(\text{IBD} = i | 2 \text{ relatives of type R are affected})}{P(2 \text{ relatives of type R are affected})}$$

for $i = 0, 1, 2$.

These probabilities are computed to be

$$z_{R0} = \alpha_{R0} \frac{1}{\lambda_R}$$
$$z_{R1} = \alpha_{R1} \frac{\lambda_O}{\lambda_R}$$
$$z_{R2} = \alpha_{R2} \frac{\lambda_M}{\lambda_R}.$$

Where, $\lambda_O$ and $\lambda_M$ are the relative risks to offsprings and monozygotic twins of an affected individuals, respectively.

To calculate the IBD sharing probabilities at a location linked to the disease locus, define $IBD_m$ and $IBD_t$ as the number of alleles shared IBD by the relative pair at the marker.

---

1If $X$ is the trait value, $Z_1$ and $Z_2$ the parental alleles, then $V_A = 2\text{Var}[E(X|Z_1)]$ and $V_D = \text{Var}[X - E(X|Z_1) - E(X|Z_2)]$, [Lange, 1997]
and trait locus, respectively. Define the parameter $\Psi = \theta^2 + (1 - \theta)^2$, $\theta$ denoting the recombination fraction between the two loci [Haseman and Elston, 1972]. The probabilities $z_{Ri}(\theta)$ at a marker locus that is recombination fraction $\theta$ from the trait locus are

$$z_{Ri}(\theta) = \frac{P(\text{IBD}_m = i | \text{2 relatives of type R are affected})}{P(\text{2 relatives of type R are affected})}$$

$$= \alpha_{Ri} \frac{1}{K(1 - K_R)} \sum_{j=0}^{2} P(\text{2 relatives of type R are affected} | \text{IBD}_t = j) \times P(\text{IBD}_t = j | \text{IBD}_m = i)$$

For R=S, i.e., for sibling pairs, the above equation leads to

$$z_{S0}(\theta) = \frac{\alpha_{S0}}{\lambda_S} [\Psi^2 + \lambda_O 2 \Psi (1 - \Psi) + \lambda_M (1 - \Psi)^2]$$

$$z_{S1}(\theta) = \frac{\alpha_{S1}}{\lambda_S} [\Psi (1 - \Psi) + \lambda_O (\Psi^2 + (1 - \Psi)^2) + \lambda_M \Psi (1 - \Psi)]$$

$$z_{S2}(\theta) = \frac{\alpha_{S2}}{\lambda_S} [(1 - \Psi)^2 + \lambda_O 2 \Psi (1 - \Psi) + \lambda_M \Psi^2].$$

The parameter $\lambda_S$ is the relative risk to the sibling of an affected individual. Similar relations are obtained for Grandparent-Grandchild pairs, Uncle (Aunt) - Nephew (Niece) pairs, Half-Sib pairs and First-Cousin pairs. Given the relation between the relative risks and $V_A$, $V_D$, and $K$, one can rewrite the IBD sharing probabilities $z_{Si}(\theta)$ in terms of $\theta$, $K$, $V_A$, and $V_D$.

The main conclusions of this article are as follows. The IBD sharing for grandparent-grandchild pairs is least affected by recombination, followed by sibs, half-sibs, uncle-nephew, and first-cousin pairs, in that order. For diseases with large $\lambda$ values and for small $\theta$ values distant relatives provide greater power. For larger $\theta$ values, grandparent-grandchild pairs are best; for small $\lambda$ values, sibs are best.

Risch [1989] described a maximum likelihood method to estimate the IBD probabilities.
Risch [1990c] extended it to a likelihood ratio test, the MLS method is described for relative pairs to assess linkage. As before, let $\alpha_i$ be the prior probability of two relatives sharing $i$ alleles IBD and let $z_i$ be the posterior probability that the two relatives share $i$ marker alleles IBD given that they are both affected, for $i = 0, 1, 2$. The subscript ‘R’ denoting the relative type is dropped in what follows. The parameters $z_i$ also depends on $\theta$. The goal is to estimate the $z_i$’s from observed marker data. For relative pair $j$, let $w_{ij}$ be the probability of the observed marker phenotypes (and of possibly other relevant relatives) given that they share $i$ marker alleles IBD, $i = 0, 1, 2$. Let $z = (z_0, z_1, z_2)$, $G_i$ denote the marker phenotype for ARP $i$, and $IBD_i$ the number of alleles shared IBD by relative pair $i$ at the marker being studied. The likelihood of the parameter $z$ for the observed data of the $j$th relative pair is

$$L_j = P_z(G_j) = \sum_{i=0}^{2} z_i w_{ij}.$$  

The likelihood for $N$ independent pairs of relatives is then given by $L = \prod_{j=1}^{N} \left( \sum_{i=0}^{2} z_i w_{ij} \right)$. The likelihood ratio can be written as

$$\Lambda = \prod_{j=1}^{N} \left( \frac{\sum_{i=0}^{2} z_i w_{ij}}{\sum_{i=0}^{2} \alpha_i w_{ij}} \right)$$  \hspace{1cm} (2.1)$$

The $\log_{10}$ of the likelihood ratio can be interpreted as a LOD score:

$$T = \log_{10} \Lambda$$  \hspace{1cm} (2.2)$$

A LOD score can be obtained by maximizing equation (2.2) with respect to $z$. By analogy to the conventional significance criterion of a LOD $> 3$ [Morton, 1955], we can apply the same criterion to $T$. Instead, we could use the distribution of $T$ and get an appropriate rejection criterion with known levels of significance Holmans [1993], Self and Liang [1987].

The parameter $z$ can be estimated by maximum likelihood, which is equivalent to maximiz-
ing the LOD score since the denominator of the LOD score is a constant. The Expectation-Maximization (EM) algorithm [Dempster et al., 1977] is used by Risch [1990c] to obtain the maximum likelihood estimates. The EM updates of \( z \) are given by the following equation

\[
\begin{aligned}
z_i^{(n+1)} &= \frac{1}{N} \sum_{j=1}^{N} \frac{z_i^{(n)} w_{ij}}{z_0^{(n)} w_{0j} + z_1^{(n)} w_{1j} + z_2^{(n)} w_{2j}} \\
\end{aligned}
\]

(2.3)

The details of the EM which are given in the appendix (B.0.5).

The expected MLS is calculated as a function of the polymorphism information content\(^2\) (PIC) for various values of the relative risk ratio, \( \lambda_R \). Two sampling strategies are studied, Scheme 1: data is available for affected relative pairs only, and Scheme 2: additional informative relatives are available. The primary conclusions of the article are that, when additional relatives are typed, they significantly increase the information about allele sharing when the markers are not highly polymorphic (PIC < 0.7, considering ASPs), i.e. Scheme 2 is better than Scheme 1. For large \( \lambda_R \) and PIC values, distant relatives are preferred while for small relative risks and PIC values sib pairs are the best.

### 2.3 MLS Under Possible Triangle Constraints

Holmans [1993] extended the MLS test and evaluated the distribution of the modified MLS statistic. The paper sets out to do three things: (1) To introduce an improvement to Risch’s method to increase power for low polymorphism markers by restricting maximization to genetically possible models only; (2) To obtain the asymptotic distributions for the test statistics, enabling criteria for a test of any size to be found; and (3) To use these asymptotic distributions to evaluate the necessity of parental genotypes. It is to be noted that though

\(^2\)The polymorphism information content (PIC) value is a measure of polymorphism introduced by Botstein et al. [1980] to describe a genetic marker’s usefulness for linkage analysis when localizing a gene involved in a rare dominant disease. If a marker has \( n \) alleles, \( i^{th} \) allele having frequency \( p_i \), under Hardy-Weinberg Equilibrium the PIC = \( 1 - \sum_{i=1}^{n} p_i^2 - 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_i p_j \).
not apparent from the original paper [Risch, 1990c] Risch did use restriction criteria [Risch, 1992], though he did not calculate the asymptotic distribution of his test statistic.

Not all IBD sharing conditional probability vectors correspond to a possible genetic model. Risch [1990b] showed that for a single locus disease model, the values of $z$ for affected sib pairs at the trait locus have the relations $z_0 = 0.25/\lambda_S$, $z_1 = 0.5 \times \lambda_O/\lambda_S$ and $z_2 = 0.25 \times \lambda_M/\lambda_S$. By expressing the relative risks in terms of population additive and dominant variances [James, 1971], it can be seen that $\lambda_O \leq \lambda_S$, with equality holding for non-dominance traits. It is also true that $\lambda_M \geq \lambda_S$. Hence the set of possible sharing probabilities for sib pairs at each disease susceptibility locus corresponds to the triangle in the $(z_0, z_1)$ plane bounded by the lines $z_0 = 0$, $z_1 = 1/2$ and $z_1 = 2z_0$. Figure 2.1 shows this triangle bounded by dashed lines. These inequalities also hold true for the sharing probabilities at a marker locus, whatever the recombination fraction $\theta$ between the trait and marker locus. The article proposes to perform the unrestricted maximization of the parameter $z$ and then to restrict it to the triangle using some rules.

Figure 2.1: Triangle constraints. The triangle with the dashed boundary denotes the values that the pair $(z_0, z_1)$ can take.

The unrestricted test statistic, $2 \ln(\Lambda) = 2 \ln \left( \frac{L(\hat{z})}{L(1/4, 1/2, 1/4)} \right)$ has an approximate
\( \chi^2 \) distribution with 2 degrees of freedom. Following Self and Liang [1987] the asymptotic distribution of of the MLS under the possible triangle constraints is a mixture of \( \chi^2_1 \) and \( \chi^2_2 \). The proportional improvement gained over the unrestricted test is greatest for low-PIC markers, and in certain cases (small effect or \( \theta = 0 \)) it can almost double the power.

The approximate asymptotic powers for 100 pairs with typed parents and 100 pairs without typed parents are compared. It is observed that pairs without typed parents perform quite well. In particular, under the simulation settings of this article, their power is always at least one-half that of the same number of pairs with parents (even for PIC as small as 0.4) - and it is often much greater, especially for high PIC.

2.4 A General Conditional Logistic Model for Affected Relative Pair Linkage Analysis

Olson [1999] proposed the Conditional Logistic model for ARP linkage analysis. The model is derived from the MLS statistic, but can include any general relative pairs, discordant relative pairs, covariates, or additional disease loci in a unified framework. The model is parameterized in terms of the logarithms of \( \lambda_M \) and \( \lambda_O \) [Risch, 1990a]. For ease of exposition let \( \lambda_M = \lambda_2 \), \( \lambda_O = \lambda_1 \) and \( 1 = \lambda_0 \), the relative risk to a person of sharing 2,1 or 0 alleles IBD at the disease locus with an affected individual.

Let \( I_m \) denote the available marker data, \( \alpha_i \) be the unconditional probability that a relative of type \( R \) shares \( i \) alleles IBD, and \( \hat{f}_{Ri} \) be the probability of sharing \( i \) alleles IBD conditional on the available marker data \( I_m \). The likelihood for one pair proposed here can be derived
from (2.1), as follows.

\[ \Lambda = \frac{\sum_{i=0}^{2} z_i w_{ij}}{\sum_{i=0}^{2} \alpha_i w_{ij}} = \frac{\sum_{i=0}^{2} z_i \hat{f}_{R_i} P(I_m|R)}{\sum_{i=0}^{2} \alpha_i \hat{f}_{R_i} P(I_m|R)} = \sum_{i=0}^{2} z_i \hat{f}_{R_i} \alpha_i. \]

At the disease locus we have the relations \( z_0 = 0.25/\lambda_S, z_1 = 0.5 \times \lambda_O/\lambda_S \) and \( z_2 = 0.25 \times \lambda_M/\lambda_S \). Also using the fact that \( \lambda_S = .25 + .5\lambda_O + .25\lambda_M \) we have

\[ \Lambda = \frac{\sum_{i=1}^{2} \hat{f}_{R_i} \lambda_i}{\sum_{i=1}^{2} \alpha_i \lambda_i} \]

Reparametrizing the model in terms of the logarithms of genetic relative-risk parameters, \( \beta_i = \ln \lambda_i \), the likelihood ratio becomes

\[ \Lambda = \frac{\sum_{i=1}^{2} \hat{f}_{R_i} \exp^{\beta_i}}{\sum_{i=1}^{2} \alpha_i \exp^{\beta_i}}. \]

The triangle constraints on \( z_i \) translates into constraints on \( \beta_i \).

Covariates may provide information about heterogeneity in linkage analysis. If we wish to condition on the value of the covariate \( x \), the likelihood ratio can be written as

\[ \Lambda = \frac{\sum_{i=1}^{2} \hat{f}_{R_i} \exp^{\beta_i + \delta_i x}}{\sum_{i=1}^{2} \alpha_i \exp^{\beta_i + \delta_i x}}. \]  

The model can also be easily extended to include Discordant Sib Pairs.

The ability to use covariates in linkage analysis helps genetic epidemiologists tackle the problem of genetic heterogeneity. In this analysis, inclusion of a covariate allows for linkage heterogeneity due to the covariate; for example, a binary covariate differentiating two
populations with different causes of disease provides more power to identify the disease genes.

Goddard et al. [2001] extended the Conditional Likelihood of Olson [1999] to a one parameter model. Addition of each covariate to (2.4) requires two additional parameters. Goddard et al. [2001] constrained the relative risks so that $\lambda_2 = 3.634\lambda_1 - 2.634$, reducing, from two to one, both the number of parameters in the basic model and the number of additional parameters needed for each added covariate. This particular constraint is derived from those given by Whittemore and Tu [1998], who showed that a min-max one-parameter ASP LOD score preserved type I error but had more power for most genetic models than did the usual two parameter model. This constraint assumes a genetic model about halfway between recessive and dominant. The LRT has a distribution which is a mixture of $\chi^2$ statistics. With the degrees of freedom depending on the number of covariates.

### 2.5 Linkage Tests for Affected Pedigree Members

Whittemore and Halpern [1994] proposed a class of non-parametric tests for linkage between a marker and a disease susceptibility gene. The tests are formed by assigning a score to each possible inheritance vector\(^3\) of the pedigree and then averaging the scores over all possible patterns compatible with the observed marker genotypes and relationships of the affected members. Different score functions give different tests. One score function, $S_{\text{pair}}(\cdot)$, is defined as the number of pairs of alleles from distinct affected pedigree members that are IBD. A second function, $S_{\text{all}}(\cdot)$, examines allele similarity across arbitrary subsets, not just pairs, of affected members, giving more weight to alleles that are shared by more

\(^{3}\)The inheritance pattern at each locus $x$ is completely described by a binary inheritance vector, $v(x) = (p_1, m_1, p_2, m_2, \ldots, p_n, m_n)$, whose coordinates describe the outcome of the paternal and maternal meiosis giving rise to the $n$ non-founders in the pedigree. Specifically, $p_i = 0$ or 1, according to whether the grandmaternal or grandpaternal allele was transmitted in the paternal meiosis giving rise to the $i^{th}$ non-founder; $m_i$ contains the corresponding information about the maternal transmission. The inheritance vector completely specifies which of the $2f$ founder alleles are inherited by each non-founder.
than two affected individuals. The resulting test is more powerful in many situations, using more information than pair based tests. Kruglyak et al. [1996] generalized this method to include both parametric and non-parametric linkage methods. In parametric linkage tests, the LOD score method is obtained with a particular score function. They also incorporated both parametric and non-parametric linkage analysis tools into the software suite GENE-HUNTER.

The original NPL tests [Kruglyak et al., 1996] are conservative when the descent information is incomplete, which is nearly always the case. The variance of the statistic is computed under the approximation that the marker is completely informative. This leads to an overestimate of the variance of the statistic. The bias is severe when the information is far from complete, in which case the procedure can be unacceptably conservative. Kong and Cox [1997] proposed a modification to the original NPL tests. Under missing data patterns, it allows for an exact calculation of likelihood ratios. The modifications of Kong and Cox [1997] significantly improved the power of NPL statistics for linkage analysis.

### 2.6 Comparison of Model Free Linkage Statistics

All model free linkage statistics are based on the premise that a set of ASPs will share more than the expected proportions of alleles at a disease susceptibility locus. There are numerous methods for testing the hypothesis that the allele sharing for a group of ASPs at a locus is $> 1/2$. The paper by Davis and Weeks [1997] aimed to answer the following question. Which method is best suited for the data at hand?

The paper evaluates 23 different model free linkage test statistics by applying them to single marker data simulated under several two-locus disease models. The statistics evaluated fall into two broad classes: (1) those that test directly for increased IBS or IBD sharing (Mean test, Proportion test, Haseman-Elston regression, NPL) and (2) those that are based on
likelihood ratio tests and that report LOD scores (MLS and various derivatives). For each of eight two-locus disease models, the paper analyzed six data sets; the first three data sets considered families with two affected siblings and 0, 1, or 2 parents genotyped. The other three data sets had four-child families with at least two affected siblings and 0, 1, or 2 parents genotyped. The authors provide false-positive rates, power of each test statistic, and overall rank by power. The powers of different test statistics are found by averaging over different disease models and family types.

The empirical false-positive rates for most of the statistics are, in general, near or slightly less than the nominal level. An important conclusion of the paper is that in terms of power, those statistics that measure IBS sharing fare much worse than those that measure IBD sharing. Because an allele that is IBS may not be IBD, there is generally less information in IBS sharing than in IBD sharing. Hence, the observed difference in the powers of IBS and IBD methods is to be expected. Also noted is the fact that, of the ten most powerful statistics, all but two (Mean test and Haseman-Elston regression) are likelihood-based tests. The likelihood based methods are generally very good at maintaining power when parental genotypes are not available. The decrease in power due to untyped parents was even lesser for larger families. Phenotypically normal individuals when used in the calculations was found to increase the power of linkage methods. For example, the Haseman-Elston regression was found to have the greatest average power when families with larger sibships were used.

The two NPL test statistics $S_{\text{all}}$ and $S_{\text{pairs}}$ were found to be extremely conservative and hence lacking in power. The reason for this is that when ambiguous IBD sharing is encountered the variance of the test statistic is an overestimate of the true variance. This is the problem that Kong and Cox [1997] aimed to address. The proposals of Kong and Cox [1997] are not incorporated in the comparison of Davis and Weeks [1997].
Choosing pairs from siblings: When more than two affected siblings are present in a sibship, all possible pairs of siblings are not statistically independent [Hodge, 1984]. Hence, some programs [Kruglyak and Lander, 1995] allow the user to select one of the three schemes: (1) use only one affected pair per family; (2) use all independent pairs; (3) use all pairs of affected siblings. In families with greater than two affected siblings, it was found that using all possible ASPs from each family lead to greater power than using either independent pairs or only one pair. Furthermore, using all possible pairs was found to be more conservative than the other two schemes. Thus, the findings of Davis and Weeks [1997] imply that using all pairs in a family is as powerful as, and more conservative than, using independent pairs or a single pair from each family. It is important to note though that this finding is contradictory to those of Greenwood and Bull [1999]. The use of all pairs can be detrimental if large sibships are in the data and may lead to underestimation of p-values in the tails (though not in the body) of the distribution. This apparent contradiction could be because of the fact that Davis and Weeks [1997] have used a maximum sibship size of four, and not all of those sibs need to be affected.
Chapter 3

Confidence Set Inference

Recent advances in genotyping technology have led to an increase in the number of genome-wide experiments being conducted. Such an explosion in the availability of genetic data demands a paradigm shift in linkage analysis. While in the past the identification of genomic regions linked to a trait was of primary importance, the much denser maps available today promise better localization.

Traditional linkage analysis involves testing for linkage, even weak linkage (the alternative hypothesis being $\theta < 1/2$). With the advent of genome scans with densely spaced markers it is possible to look for tight linkage, not just linkage. Lin et al. [2001] proposed Confidence Set Inference (CSI), a statistical framework for identifying tight linkage (with, say $\leq 5$ cM between the marker and disease loci) with backcross data. The primary objective of the method is to provide a confidence set of regions on the genome which house the disease gene with a predetermined probability. In addition, the method eliminates the need to correct for “multiple testing” (under some assumptions). The problem of multiple testing is that of controlling the experiment-wise error rate to the $\alpha$ level (type I error of a single test) does not control the family-wise error rate at $\alpha$ (the probability that at least one test is rejected given that all null hypothesis are true).
CSI provides a promising framework for constructing precise confidence sets for disease gene locations. Lin [2002] reformulated two popular model free linkage tests for affected sibling pair (ASP) linkage, the Mean and Proportion test, in the CSI framework. Papachristou and Lin [2005a] extended the framework to test markers with incomplete IBD sharing information. To handle the dense Single Nucleotide Polymorphism (SNP) maps available today, Papachristou and Lin [2006a] proposed two multi-point CSI methods. However, because the distribution of the test statistic needs to be evaluated at the disease locus, the procedure requires knowledge of some disease model related parameters. While these parameters may be estimated from population epidemiologic studies for single-locus disease models, their estimates for complex diseases may be suspect. To tide over this problem, Papachristou and Lin [2006b] proposed a two-step CSI procedure that estimates these parameters from a coarse map of markers (e.g., a microsatellite map), and uses them in the second step on a fine map of markers (e.g., a SNP map) to construct CSI confidence sets.

In addition to CSI, there are other methods of constructing confidence sets for disease gene locations, namely, the LOD support intervals [Ott, 1999], a Generalized Estimating Equations approach [Liang et al., 2001a], Bootstrap methods [Papachristou and Lin, 2006c], and some less studied approaches (e.g., Hauser et al. [1996], Hössjer [2003]). In a comparison of these methods, the two-step CSI approach was shown to perform well [Papachristou and Lin, 2006c]. In this Chapter, CSI and its various extensions are reviewed.

### 3.1 CSI for Model Based Linkage Analysis

Lin et al. [2001] proposed the CSI approach to construct random confidence sets that include the disease gene location with a predetermined probability. A further theoretical advantage of CSI is that rewriting the null hypothesis of traditional linkage this way circumvents the need to perform multiple testing corrections for the number of markers being
3.1.1 Multiple Testing in Linkage Analysis

Linkage analysis involve performing the following hypothesis test for each marker in a study

$$H_{0m}^{(\text{trad})}: \theta_m = 1/2 \quad \text{against} \quad H_{am}^{(\text{trad})}: \theta_m < 1/2,$$

(3.1)

where $\theta_m$ is the unknown recombination fraction between marker $m$ and the disease locus. The markers at which the test statistics take values more adverse than some predetermined cut-off are declared linked to the disease locus. Recent developments in genetics have led to anywhere between a few hundreds to thousands of dense markers being available across the whole genome. Such a high number of (often dependent) tests leads to the problem of multiple testing. Following theory from sequential testing, Morton [1955] showed that a LOD score of 3.0 may be used as a cutoff for declaration of linkage in mapping genetic traits by means of a relatively small number of markers. However this recommendation was not for current genome scans which involve thousands of markers and is generally regarded as not being stringent enough to avoid too many false positives. In light of the potential for inflated type I error when multiple markers are tested by the LOD method, several different cutoffs were proposed, based on diverse arguments, including exact calculation, Bonferroni correction, and Bayesian formulation.

Several cutoff values, for different scenarios, were given by Lander and Kruglyak [1995], the most often quoted being 3.6. This has lead to an ongoing debate about the merits of such general cut-offs with regard to their actual coverage [Witte et al., 1996, Curtis, 1996, Rao, 1998, Morton, 1998, Sawcer et al., 1997, Zhao et al., 1999]. In particular, via a whole-genome simulation study, Sawcer et al. [1997] showed that the use of a LOD score of 3.2 ensures a genomewide significance level of 5%, leading them to argue in support of
the LOD-score-3.0 criterion. Zhao et al. [1999] conducted Monte Carlo experiments and opined that under typical settings a cut-off of 3.6 is conservative while a cut-off 3.0 leads to more false positives than expected. The central issue of this debate is how to control the type I error to an acceptable rate genomewide and yet be able to detect signals for linkage when multiple tests are performed. Other measures of genomewide significance have been proposed including the false-discovery rate [Drigalenko and Elston, 1997], false positive predictive error and the reliability index [Morton, 1998], and expected number of false positive errors [Zhao et al., 1999].

The CSI method is a statistical procedure for finding tightly linked markers. It aims to identify only markers that are within a small distance (say within $d_0$ cM) of the disease locus. The idea is to construct a confidence set of markers, with coverage probability at a predetermined level, $p$. In other words, the method’s output is a set of markers $A$, every disease gene is within a distance $d_0$ cM of one of the markers in $A$, with a probability $p$, where $d_0$ is chosen such that asymptotically a unique marker is within this distance of the disease gene. With the new hypothesis, there is only a small number (possibly one, under ideal situations) of null hypotheses that are known to be true a priori, and this means that the procedure does not need to be corrected for the number of markers being tested.

3.1.2 Construction of a Confidence set

Suppose that we have $M$ markers along the chromosome and the largest distance between any two consecutive markers is $d$, define $d_0 = d/2$. We know that the disease locus has at least one marker within $d_0$ cM of it. If all the markers are equally spaced and the disease locus does not lie exactly between two markers, then there is a unique marker, $m^*$, that is within $d_0$ distance of the disease gene. The goal is to find $A$ such that $P(m^* \in A) \geq p$. Such a set of markers is referred to as a “confidence set” with a coverage probability $p$. 

29
The new hypothesis is formulated as

\[ H_{0m}^{(new)} : d_m \leq d_0 \text{ vs. } H_{am}^{(new)} : d_m > d_0, \]

where \( d_m \) is the true but unknown distance between the disease locus and marker \( m \). Rewriting the hypothesis in terms of the recombination fractions, we have

\[ H_{0m}^{(new)} : \theta_m \leq \theta_0 \text{ vs. } H_{am}^{(new)} : \theta_m > \theta_0, \tag{3.2} \]

where \( \theta_m \) is the recombination fraction between marker \( m \) and the disease locus, and \( \theta_0 \) the equivalent of \( d_0 \) in recombination fractions. The important difference in writing the hypothesis in this form is that, now, the null hypothesis is that of tight linkage. Also, exactly one null hypotheses is known to be true. The generalized likelihood-ratio statistic is

\[
\lambda_m = \sup_{\theta_m \in [0, \theta_0]} \frac{L(\theta_m)}{\sup_{\theta_m \in [0,1/2]} L(\theta_m)} = \begin{cases} 1 & \text{if } \hat{\theta}_m \leq \theta_0 \\ \frac{L(\theta_0)}{L(\hat{\theta}_m)} & \text{if } \hat{\theta}_m > \theta_0 \end{cases},
\]

where \( \hat{\theta}_m \) is the maximum likelihood estimate (MLE) of \( \theta_m \) obtained by maximizing the likelihood function \( L(\theta_m) \). Since a small value of \( \lambda_m \) provides evidence against the null, the rejection region is given by \( c_\alpha (\leq 1) \) such that

\[
\sup_{\theta_m \leq \theta_0} P_{\theta_m} \left( \frac{L(\theta_0)}{L(\hat{\theta}_m)} < c_\alpha \right) = \alpha,
\]

where \( \alpha = 1 - p \) is the type I error of our test. Then \( A = \{ m : \lambda_m \geq c_\alpha \} \) is a confidence set with coverage probability \( P(m^* \in A) = P_{\theta_m^*}(\lambda_m^* \geq c_\alpha) \geq 1 - \alpha = p. \)

To verify whether the above formulation of confidence sets will be of use in genetics, the authors construct confidence sets of markers for two commonly used family data types. The data has \( n \) families each with two siblings. The first type of families is the phase...
known (PK)\textsuperscript{1} double backcross and the second, the phase unknown (PU) double backcross. In both types of families one of the parents is doubly homozygous (not informative for linkage) while the other parent is doubly heterozygous.

**Phase Known Sib-Pair data**

The PK families allow for exact evaluation of the number of recombination events. Let $S_m = \sum_{i=1}^{n} X_i$ be the number of recombination events out of a maximum of $2n$ meiosis, where $X_i$ is the number of recombination events in sib-pair (SP) $i$. Then the likelihood for $\theta_m$ is

$$L(\theta_m | x_1, \ldots, x_n) = \prod_{i=1}^{n} \left( \frac{2}{x_i} \right)^{x_i} (1 - \theta_m)^{2-x_i} \propto \theta_m^{s_m} (1 - \theta_m)^{2n-s_m},$$

where $s_m$ is a realization of $S_m$. So, $S_m$ is a sufficient statistic for $\theta_m$ and the likelihood ratio test amounts to rejecting the null hypothesis when $S_m$ is large, since a larger values of $\theta$ leads to greater chance of recombination.

$$\alpha = \sup_{\theta_m \leq \theta_0} P_{\theta_m}(S_m > c_\alpha) = P_{\theta_0}(S_m > c_\alpha) = \sum_{s_m > c_\alpha} \left( \frac{2n}{s_m} \right) \theta_0^{s_m} (1 - \theta_0)^{2n-s_m} \approx 1 - \Phi \left[ \frac{c_\alpha - 2n\theta_0}{\sqrt{2n\theta_0(1 - \theta_0)}} \right]$$

where $\Phi$ is the cumulative distribution function of the standard normal variable. The above equation can be solved to get the value of $c_\alpha$, given by the equation

$$c_\alpha = 2n\theta_0 + \Phi^{-1}(1 - \alpha) \sqrt{2n\theta_0(1 - \theta_0)}.$$

Similar computations lead to the threshold required to reject the CSI null for phase un-

\textsuperscript{1}The situation when it is known whether the AB/ab parent’s two chromosomes are actually AB/ab or Ab/ab.
known double backcross data.

### 3.1.3 Results and Discussion

The authors compared results from the LOD method with a cut-off of 3.6 with a 99% CSI confidence sets. It was observed that the CSI confidence sets were much more successful at including only markers closer to the disease locus. Another important distinction is that, as the sample size increased, the CSI confidence sets became more precise while the number of markers with LODs greater than 3.6 increased. This is to be expected since as the sample size increases, in traditional linkage one has more power to reject the null hypothesis $\theta_m = 1/2$, when $\theta_m < 1/2$ even though close to 1/2. The LOD method is designed to pick up any linkage, even weak linkage.

There are two main advantages of the CSI approach. CSI provides confidence sets that can identify sufficiently localized genomic regions for linkage, so that the need for efforts at further localization is greatly reduced after an initial screen. Secondly, for a single locus disease model, there is no need to be concerned with the multiplicity-adjustment problem, which is encountered when screening a large number of markers.

The general principle of multiplicity adjustment is that multiplicity needs be adjusted only to the extent that the null hypotheses being tested may be true simultaneously. Consider the example of bioequivalence studies. Two drugs are said to be bioequivalent if they are absorbed into the blood and become available at the same rate and concentration [Berger and Hsu, 1996]. To prove equivalence use the null hypothesis that the absorption of the new drug is considerably greater or lesser than that of the existing drug. One can conduct two one-sided tests simultaneously to test the hypothesis; however, no multiplicity adjustment is needed, because it is impossible to deliver both too much and too little of the test drug at the same time. This implies that, when a genetic disorder involves a single disease gene, no
multiplicity adjustment is needed, no matter how many markers throughout the genome are tested one by one. In the description of the method, for ease of presentation it is assumed that there exists a unique marker within a distance $d_0$ of the disease gene. This assumption is not necessary and may be overcome by using a Bonferroni type correction for the number of disease genes involved, a number much smaller than the number of tests performed.

### 3.2 CSI for Affected Sibling Pair Linkage Analysis

Lin [2002] extended CSI to Affected Sibling Pair (ASP) linkage analysis. Two well known non-parametric linkage statistics, the Two-IBD test, or Proportion test, and the Mean test are reformulated to test the CSI hypothesis. The simulation results show that the fundamental advantages of the approach in model based linkage analysis is retained when the method is extended to model free linkage analysis. As in parametric linkage analysis, the following hypothesis is tested

$$H_{0m}: \theta_m \leq \theta_0 \text{ vs. } H_{am}: \theta_m > \theta_0,$$

where $\theta_m$ is the recombination fraction between marker $m$ and the disease locus and $\theta_0$ is chosen to control the maximum number of null hypothesis, a priori, that can be true.

#### 3.2.1 Two Test Statistics and their Properties

The Two-IBD and the Mean test are reformulated to test the CSI hypotheses [Lin, 2002] and are detailed in the following sections.
Two-IBD Test

The Two-IBD test, also called Proportion test, is given by the test statistic $T_m$ at marker $m$, which is the proportion of ASPs who share two alleles IBD. The traditional hypothesis test, given by

$$H^{(\text{trad})}_{0m} : z_2(\theta) = 1/4 \text{ versus } H^{(\text{trad})}_{am} : z_2(\theta) > 1/4,$$

is replaced by the new set of CSI hypotheses

$$H^{(\text{new})}_{0m} : z_2(\theta) \geq z_2(\theta_0) \text{ versus } H^{(\text{new})}_{am} : z_2(\theta) < z_2(\theta_0).$$

Where $z_2(\theta)$ is defined as $z_i(\theta) = P(mIBD = i | \text{Affected Sibling Pair})$ [Risch, 1990b]. The inequalities in the hypotheses are natural since, as shown in Lemma B.0.6, $z_2(\theta)$ is decreasing in $\theta$. Under no linkage, two siblings will share $(0,1,2)$ alleles IBD with probability $(1/4,1/2,1/4)$. The IBD sharing probability for two siblings at a location which is $\theta$ recombination fraction from the disease locus, $z_i(\theta)$, may be written as [Lin, 2002],

$$z_0 = \frac{K^2 + (1 - \Psi)V_A + (1 - \Psi)^2V_D}{4K^2 + 2V_A + V_D},$$
$$z_1 = \frac{2K^2 + V_A + 2\Psi(1 - \Psi)V_D}{4K^2 + 2V_A + V_D},$$
$$z_2 = \frac{K^2 + \Psi V_A + \Psi^2V_D}{4K^2 + 2V_A + V_D}.$$

Where $K$ is the population prevalence, $V_A$ the additive genetic variance and $V_D$ the dominant genetic variance and $\Psi = \theta^2 + (1 - \theta)^2$.

For large enough sample size $n$ (say, $nz_2(\theta_m) \geq 10$), $T_m$ is approximately normally distributed. Under the traditional hypothesis the critical value for a level $\alpha$ test is given by

$$\frac{1}{4} + \Phi^{-1}(1 - \alpha)\sqrt{3/16n},$$
where $\Phi$ is the cumulative distribution function of a normal vari-
ate. By Lemma B.0.6, if $\theta_m < 1/2$, then $z_2(\theta) > 1/4$ and consequently,

$$\lim_{n \to \infty} P(T_m > 1/4 | \theta_m) = 1$$ \hspace{1cm} (3.3)$$

which follows by the weak law of large numbers (WLLN). However,

$$\lim_{n \to \infty} \left( \frac{1}{4} + \Phi^{-1}(1-\alpha) \sqrt{\frac{3}{16n}} \right) = \frac{1}{4},$$ \hspace{1cm} (3.4)$$

Equations (3.3) and (3.4) together imply that under the traditional formulation, as data accumulate, the test statistic for a marker on the same chromosome as the disease locus will exceed any given threshold for significance, leading to declaration of linkage. Under the new formulation of the hypothesis (3.2), we reject $H_{0m}^{(n\text{ew})}$ if $T_m < C_\alpha$, where $C_\alpha$ is obtained from

$$\sup_{\theta_m \leq \theta_0} P(T_m < C_\alpha | \theta_m) = P(T_m < C_\alpha | \theta_0) = \alpha,$$

since a larger recombination fraction between marker and disease leads to a smaller probability of having two marker alleles IBD. Using a normal approximation one can get the critical value given in Table 3.1. Each marker for which the the null hypothesis has been rejected is included in the confidence set $A$. That is, $A = \{m : T_m \geq C_\alpha\}$. If $\theta_m > \theta_0$, then $z_2(\theta_m) < z_2(\theta_0)$ and the critical value is $z_2(\theta_0) + \Phi^{-1}(\alpha) \sqrt{\frac{z_2^2(\theta_0)(1-z_2^2(\theta_0))}{n}}$. By an application of the WLLN we have,

$$\lim_{n \to \infty} P(T_m < C_\alpha | \theta_m) = 1$$

meaning that with a large enough sample size, any marker located farther than $\theta_0$ from the disease gene will not be included in the confidence set.

---

2For a proof see Lemma B.0.7
Table 3.1: Two-IBD test statistic properties

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$\sum_{i=1}^{n} I(mIBD_i = 2)/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$z_2(\theta)$</td>
</tr>
<tr>
<td>Variance</td>
<td>$z_2(\theta)(1 - z_2(\theta))/n$</td>
</tr>
<tr>
<td>Critical Value</td>
<td>$z_2(\theta_0) + \Phi^{-1}(\alpha)/z_2(\theta_0)(1 - z_2(\theta_0))/n$</td>
</tr>
<tr>
<td>Power ($\theta &gt; \theta_0$)</td>
<td>$\Phi(C_\alpha - z_2(\theta)\sqrt{z_2(\theta)(1 - z_2(\theta))/n})$</td>
</tr>
<tr>
<td>Sample Size (Power $= q$)</td>
<td>$\lceil\left(\left(\Phi^{-1}(q)\sqrt{z_2(\theta)(1 - z_2(\theta))} - \Phi^{-1}(\alpha)\sqrt{z_2(\theta_0)(1 - z_2(\theta_0))}\right)\right)^2\rceil$</td>
</tr>
</tbody>
</table>

Mean Test

The Mean test for the CSI hypothesis is given by the test statistic $M_m$, which is the average number of marker alleles shared IBD for $n$ ASPs at marker $m$. As for the Two-IBD test; it can be shown that traditional hypothesis testing leads to a situation where as data accumulates any marker that is even loosely linked to the disease will be rejected with probability one. Also, $E_{\theta_m}(M_m)$ is decreasing in $\theta$. Hence, it is natural to reject $H_{0_m}^{(new)}$ (3.2) when $M_m < C_\alpha$, where $C_\alpha$ is such that

$$\sup_{\theta_m \leq \theta_0} P(M_m < C_\alpha | \theta_m) = P(M_m < C_\alpha | \theta_0) = \alpha. \quad (3.5)$$

The stochastic ordering in the first equality of (3.5) is not trivial, for a proof see Appendix A. The confidence set is given by $A = \{m : M_m \geq C_\alpha\}$.

3.2.2 Analysis and Results

An extensive simulation study is conducted to evaluate this method. The underlying disease model is assumed to be monogenic with two alleles and is characterized by $\{K, V_A, V_D\}$. Three disease settings are considered. The first disease model has a heritability of 71% with a recessive mode of inheritance. Under setting two, the disease has a heritability of 50%, with an additive mode of inheritance. The disease locus will be harder to detect in
this setting than in the first one. The third disease has a heritability of 14% with a dominant mode of inheritance. This setting is difficult to detect since now only 14% of the variance of the trait is accounted for by the genetic locus.

**Influence of Sample Size on Confidence Set:** Data is simulated on a disease chromosome: 30 equally spaced markers with a 10 cM spacing. The disease locus is simulated 1 cM from marker 15, between markers 15 and 16. Confidence sets with 95% coverage probability, based on both the Two-IBD test and the Mean test, are constructed for several data sets, comprising of 100, 250, 500 and 750 ASPs. A total of 1000 replications are performed for each data set. The results show a clear trend, as the number of sib pairs increases the confidence sets get smaller. The results also show that it is much easier to detect the disease locus in setting one than in the other settings and it is most difficult to detect the disease locus under setting three, but even in this situation the procedure was able to locate a disease locus up to a 30 cM region using the mean test on 750 pairs.

**Comparison of Two-IBD Test and Mean Test:** The two tests conducted under the new formulation are compared. The power of the test is set to be 80% for the alternative corresponding to a genetic distance of 10 cM. Results for the whole space of \( \{K, V_A, V_D\} \) are obtained. It is observed that the Mean test requires a smaller sample size than the Two-IBD test, especially for traits with a small prevalence. As population prevalence increases, the Two-IBD test performs at least as well as the Mean test for traits that have small additive variance. The results in this paper agree with the conclusions of Blackwelder and Elston [1985] and others that are based on the traditional formulation of the hypothesis.
3.2.3 Discussion

The traditional formulation of the null hypothesis for linkage analysis requires a necessary but difficult multiplicity adjustment in order to correctly control the genome-wide type I error rate. One can avoid this by reformulating the null and alternative hypotheses as done in this paper. Also, the confidence set obtained from this method helps localize sufficiently the results of an initial genome scan. One important obstacle to the applicability of the proposed method is that it is assumes that the markers are completely polymorphic. The situation of markers with incomplete polymorphism is explored in Papachristou and Lin [2005a], reviewed next.

3.3 CSI for Markers with Incomplete Polymorphism

Lin [2002] described a CSI procedure for construction of a confidence sets for location of the disease gene. A serious drawback of the method when applied to real data is the assumption that the markers used are completely polymorphic. Papachristou and Lin [2005a] have extended CSI to handle markers that are not 100% polymorphic. A general form of the mean test statistic is given by Holmans [1998] (implemented in S.A.G.E. [2006]), in which the indicator functions of a sib pair sharing one or two alleles IBD at the marker locus is replaced by the corresponding conditional probabilities given the marker genotypes. The proposed CSI method uses this version of the Mean test statistic. Simulations are performed to assess the performance of the method under several factors, including the number of ASPs, the heterozygosity of the marker and the availability of parental genotypes at the marker loci. The performance of the approach is also studied when the values of the risk characteristics used in the procedure are different from the true values. The results are compared to linkage analysis performed via Non-Parametric Linkage Analysis on ASPs using GENEHUNTER [Kruglyak et al., 1996, Kruglyak and Lander, 1998].
For markers with incomplete IBD sharing information, the Mean test statistic at a locus \( m \), \( M_m \), based on data from \( n \) nuclear families each with two affected sibs, is defined as

\[
M_m = \frac{\sum_{i=1}^{n} \left( P(m|IBD_i = 1|G_i^m) + 2P(m|IBD_i = 2|G_i^m) \right)}{n},
\]

where \( G_i^m \) is the observed genotype configuration of the \( i^{th} \) family at marker \( m \). Three types of data are explored, T0 when \( G_i \) consists only of the genotypes of the ASPs, T1 when it consists of the genotypes of the ASPs and one parent and T2 when genotypes of both parents are available along with the ASPs.

### 3.3.1 Models, Factors, Levels and Results

The same three models used by Lin [2002] are analyzed here. The genetic map consists of two chromosomes, each with 30 markers equi-distant from each other at 10 cM. The disease causing locus is between markers 15 and 16, 1 cM away from the former, on the first chromosome. The second chromosome contains no disease gene. The markers are assumed to be in linkage equilibrium and \( \theta_0 \) is set to be the recombination fraction corresponding to a genetic distance of 5 cM. 95% confidence sets are obtained under three models for various levels: the number of ASPs in the sample (100, 250, 500 or 750), the number of equally likely alleles at the locus (2, 5 or 8) and the type of data (T0, T1 or T2). A 1000 replicates are performed for each setting.

The number of markers included in the confidence set are similar to those obtained by Lin [2002], setting one being the easiest to detect and setting three the hardest. As would be expected, as the number of alleles at each marker increases, the size of the confidence set decreases to the situation where all markers are completely polymorphic. Increase in the number of ASPs leads to smaller confidence sets and an increase in the power. For models two and three, use of parental genotypes also leads to an increase in the power. However,
this is not the case in model one.

The mean and variance of the statistics depends on the population relative risk parameters $\lambda_0$ and $\lambda_S$, the authors explore the situation when these parameters are misspecified. They conclude that these perturbations can lead to a substantial increase in the type I error when the estimates are off by 30%, especially for diseases that have a small genetic component. On the other hand, inaccurate estimates of risk ratios have a small effect for models with a strong genetic component, or only moderate if the estimates are off by only 10%.

### 3.4 Multi-point Confidence Set Inference

The CSI procedure as incorporated in Papachristou and Lin [2005a] has two obvious disadvantages. First, it is a non-directional procedure that makes no distinction between loci symmetrically located about a marker. The confidence set thus constructed includes a region of length $2d_0$ for each marker at which the null hypothesis is not rejected. This leads to wider than desired confidence sets and overcoverage of single-point methods. The second major disadvantage of single-point CSI is the fact that it is not suitable for SNP markers, which are not very informative on their own, but can provide a wealth of information about IBD sharing when used collectively in a multi-point fashion. With the reduction in the cost of such high throughput SNP arrays it is important to be able to use them in linkage analysis. These problems can be overcome by performing a full multi-point test, that incorporates information from all markers on a chromosome to infer IBD sharing at each location [Papachristou and Lin, 2005b, 2006a]. In addition, such an approach overcomes the low informativity of individual SNPs.

In addition to proposing multi-point CSI methods, Papachristou and Lin [2006a] investigate the relative merits of SNP versus microsatellite (MS) maps for disease gene localization. Two multi-point methods for confidence set inference are proposed. A third method is
proposed that makes use of one marker at a time, and is computationally more efficient than the multi-point methods. In addition to marker type, the authors explore the effect of map density and marker heterozygosity on disease gene localization through simulations. The results show that the use of dense SNPs can lead to shorter confidence intervals for disease gene locations than MS markers.

3.4.1 Methods

The hypothesis in (3.2) does not make a distinction between locations to the left and to the right of a marker, thus the resulting confidence sets are wider than desired. To overcome this problem, Papachristou and Lin [2006a] proposed testing the following hypothesis using multi-point marker genotype information. At each location $\tau$ on the genome, test

$$H_{0\tau} : \tau = \tau^* \text{ versus } H_{A\tau} : \tau \neq \tau^*, \quad (3.6)$$

where $\tau^*$ is the true but unknown disease location. A $(1 - \alpha)$ confidence set of the disease location is $R = \{\tau : H_{0\tau} \text{ not rejected at level } \alpha\}$. Presumably, any traditional linkage test statistic could be reformulated to test the CSI nulls (3.2) and (3.6). An important step in conducting such a test is to derive the distribution of the test statistics under the CSI null of tight linkage, which requires knowledge of certain disease model related parameters.

CSI-Mean

Papachristou and Lin [2006a] reformulated the Mean test statistic to test the CSI hypothesis, at genomic location $\delta$,

$$\bar{M}_\delta = \frac{\sum_{j=1}^n \sum_{i=0}^{2} iP(\delta | BD_j = i | G_j)}{n} = \frac{\sum_{j=1}^n M_\delta(G_j)}{n}, \quad (3.7)$$
where $\delta IBD_j$ is the number of alleles shared IBD by the $j$th ASP at location $\delta$, $G_j$ denotes the appropriate family marker data for the $j$th ASP, $j = 1, \cdots, n$, and $\mathcal{M}_\delta(G_j)$ is the average number of alleles shared IBD at locus $\delta$ by the $j$th ASP given $G_j$. Let $B$ denote the event that both siblings are affected. The standardized Mean statistic is

$$t_\delta = \frac{\mathcal{M}_\delta - E(\mathcal{M}_\delta|\tau, B)}{\sqrt{\text{Var}(\mathcal{M}_\delta|\tau, B)}} = \frac{\mathcal{M}_\delta - \mu_\delta(\tau, B)}{\sigma_\delta(\tau, B)/\sqrt{n}}.$$  \hspace{1cm} (3.8)

The mean ($\mu_\delta(\tau, B)$) and the variance ($\sigma_\delta^2(\tau, B)$) are calculated under the null hypothesis that the trait locus is located at $\tau$. Papachristou and Lin [2006a] defined two multi-point variants, one of which is CSI-V3, where $\delta = \tau$ (henceforth referred to as CSI-Mean-V3). However, the IBD sharing at locations between markers cannot always be determined with much precision, hence they proposed another multi-point variant that only used IBD sharing information at markers, i.e., $\delta = m(\tau)$, where $m(\tau)$ is the marker closest to $\tau$ (we shall refer to this method as CSI-Mean-V2). A further test statistic, CSI-V1, is proposed that takes $\delta$ to be $m(\tau)$, but uses IBD information at $m(\tau)$ as inferred from the genotype at $m(\tau)$, which makes CSI-V1 a single-point method. Asymptotically, the test statistic $t_\delta$ follows the standard normal distribution under the null hypothesis. A two-sided test was suggested because the value of the test statistic at the true disease location (if different from a marker) may be lesser than the that at a marker that is close to it for lack of IBD information at the location [Papachristou and Lin, 2006a, Liang et al., 2001a].

The computation of the confidence set requires knowledge of the parameters $\mu_\delta(\tau, B)$ and $\sigma_\delta(\tau, B)$. Assuming that the $\mathcal{M}_\delta(G_j)$ are independent and identically distributed,

$$\mu_\delta(\tau, B) = E(\mathcal{M}_\delta(G_j)|\tau, B) = \sum_{G_j} \mathcal{M}_\delta(G_j)P(G_j|\tau, B).$$  \hspace{1cm} (3.9)

The variance can be written as a similar summation. Since the summation is over all possible multilocus genotypes, the computational burden grows exponentially with the number.
of markers on the map and soon becomes intractable. These two parameters depend on the IBD sharing distribution at the trait locus by an ASP, the type of pertinent relatives genotyped in addition to the siblings, and the number and heterozygosity of the markers considered. Note that \( \sum_{i=0}^{2} iP(\delta IBD_j = i|\tau, B) \neq \mu_\delta(\tau, B) \), because of incomplete IBD information. To overcome the computational difficulty, Papachristou and Lin [2006a] proposed a Monte Carlo algorithm to estimate \( \mu_\delta(\tau, B) \) and \( \sigma_\delta(\tau, B) \), recommending 6000 Monte Carlo samples per null hypothesis.

### 3.4.2 Results

To compare the relative merits of SNP versus MS maps, the authors simulated SNP maps with varying densities (every 0.25, 0.5, 0.75, and 1 cM), and MS maps with varying densities (5, 7.5, and 10 cM) on the chromosome of interest. Each SNP was assumed to have a minimum allele frequency of 0.2, 0.3, or 0.5, and the MS markers had 4, 5, or 8 equally frequent alleles. The disease chromosome was 122.5 cM long with the disease locus at 61.2 cM. The three disease models in Lin [2002] were simulated. Also explored was the effect of having zero (T0) or two (T2) parents genotyped.

The effect of marker type and density on the lengths of the constructed CSI intervals is shown. All confidence sets were found to have close to the nominal coverage of 95%. It is observed that the multi-point variants lead to considerably shorter confidence sets for the SNPs, the lack of heterozygosity being compensated for by the high density. For the MS markers, CSI-V1, which is a single point method (relying on marker heterozygosity) performs well, and CSI-V3, which uses IBD estimates at inter-marker locations, performs marginally worse than the other two variants. The multi-point methods give almost identical results for dense SNP maps. In the study of the effect of levels of marker heterozygosity, it was found that the lengths of CSI intervals reduced as the markers became more heterozygous. It was also found that, as the sample size increases, the average lengths of
the confidence sets becomes smaller while maintaining coverage. This agrees well with the fact that, as the sample size goes to infinity, the confidence set will only include the true marker location.

3.4.3 Discussion

The computation of the mean ($\mu_\delta(\tau, B)$) and the variance ($\sigma^2_\delta(\tau, B)$) of the test statistic under the null hypothesis requires knowledge of some disease model related parameters. In particular, these parameters can be the IBD sharing probabilities for two siblings at the underlying disease locus. For monogenic disease models, these can be derived from the population relative risks. The authors show ways of estimating the marginal IBD sharing probabilities for a two locus model with known interacting etiology (following the work of Risch [1990a,b]). However, complex ascertainment schemes in genetic epidemiological studies can lead to biases in these estimates [Cordell and Olson, 2000, Olson and Cordell, 2000, Zou and Zhao, 2004]. In addition, the number of interacting loci and the manner in which they interact may not be known precisely. Because of these shortcomings, Papachristou and Lin [2006b] proposed a two-step CSI approach that estimates the required disease parameters from the data at hand in the first step and uses these estimates in step two to construct the CSI intervals. This paper will be reviewed next.

In conclusion, the paper proposes two multi-point and one single-point CSI method that are capable of sufficiently localizing the underlying disease gene location, and provide confidence sets for the disease gene location with the correct confidence. Using a multi-point method means that information can be efficiently aggregated from nearby SNPs in a dense SNP chip. The use of dense SNP arrays is becoming popular and the paper shows that there are immense advantages to using them in linkage studies.
3.5 Two-Step CSI Procedure

For the derivation of confidence regions, the CSI procedure needs the distribution of allele sharing IBD by two siblings at the trait locus. Theoretically, one could estimate these probabilities using population disease characteristics, provided the number of genes contributing to the disease and the mode of interaction between them is known. Unfortunately, such information is seldom known. And, even if available, for complex diseases the estimates of these parameters will be biased. To avoid using population risk characteristics, Papachristou and Lin [2006b] proposed a two-step CSI procedure. Traditionally, gene mapping is often pursued via a linkage genome scan with microsatellite markers, and then the preliminary linkage signals are followed up using a denser, usually SNP, map. It is this practice that motivated the two-step CSI procedures, which has the following steps.

**Step I: Identification of disease region and estimation of disease parameters:** In this phase of the analysis, based on the MS data, perform a whole genome scan using the non-parametric linkage statistic (KAC-score) introduced by Kong and Cox [1997]. The KAC-score was chosen because it has been shown to perform well compared to other non-parametric linkage methods. For each region that the maximum KAC-score exceeds a preset threshold, the maximum likelihood estimators of the IBD sharing probabilities at the trait locus is computed [Risch, 1990c, Holmans, 1993]. This estimation assumes that there is only trait locus in the chromosome where a linkage signal is detected. The maximization of the likelihood is performed using a modified EM algorithm to accommodate the triangle constraints and provides consistent estimates of the true IBD sharing probabilities at the trait locus [Cordell, 2004]. The value of the threshold was chosen to be 2.33, this cut-off is often used in linkage studies to flag suggestive signals [Kruglyak et al., 1996], which is then followed by additional analysis in the region.
Step II: CSI of disease loci: The second stage of the analysis focuses on those genomic regions that give promising results in the initial scan from step one. For each of these regions, obtain genotypes for a number of SNPs located within a prespecified distance, say 25 cM, from the position where the maximum KAC-score occurs in the preliminary analysis. Using these SNP data and the IBD estimates obtained from the first phase, apply a multipoint CSI procedure [Papachristou and Lin, 2006a] to derive a confidence interval for the location of the disease causing gene. Since, the analysis is geared towards SNPs, a multipoint CSI method must be used and for dense SNPs CSI-V2 and CSI-V3 were shown to have very similar results. Hence, the authors used CSI-V3 to construct confidence sets in step two. Location by location testing can lead to jumps in the value of the test statistic, so to smooth them out a smoothing spline was used. Then, based on the smoothed version of the test statistic, the confidence set is constructed.

3.5.1 Results

Two-locus disease models, where the two loci segregate independently, are simulated. Four disease models are simulated. These belong to the twelve models studied by Knapp et al. [1994b], and are the only ones whose penetrances are not symmetric with respect to the two trait contributing loci. Two disease chromosomes are simulated of lengths 150 and 100 cM. Both are simulated to have a disease locus close to the mid-point of the chromosome and have SNP (every 0.5 cM) and MS (every 7.5 cM) markers genotyped. For sample sizes of 250 and above the two-step CSI method is able to provide precise confidence sets with correct coverage probabilities.

The CSI procedure using the Mean test estimates the mean and variance of the test statistic using Monte Carlo simulations. An assumption required for this computation is that the missing pattern in the families is constant, for example, sibs and parents genotyped, or sibs genotyped but all parents missing. However, in real genetic studies, due to reasons
including genotyping errors or inconclusive results of the assay used in the genotyping, we usually have a large number of individuals partially genotyped. So, the authors explore the effect of low levels of missing information on the resulting CSI confidence region. It was observed that the density of the genetic map (spacing of 0.5 cM) was able to compensate for the missing genotypes. It is surprising, though, that even with a relatively high amount of missing genotypes (up to 30%), the effect on the confidence region is negligible, demonstrating a high robustness of the method in this regard.

The proposed method is applied to the Arthritis Research Campaign National Repository data which consist of 157 nuclear families with two to six children affected by Rheumatoid Arthritis (RA). CSI confidence sets were constructed for the HLA-DRB1 locus on chromosome 6, a known susceptibility locus for RA. The 95% and 99% CSI confidence sets along with the 1-LOD and 3-LOD drop intervals are constructed. The lengths of the 95 and 99% confidence intervals were 6.7 and 20.4 cM, respectively, which compare favorably to those of 1-LOD and 3-LOD support intervals (8 and 31 cM, respectively). Note that 1-LOD and 3-LOD support intervals are frequently interpreted as 95 and 99% confidence intervals. However, theoretical results show that their coverage probability may be significantly lower, especially when the genetic map used in the analysis is extremely dense [Dupuis and Siegmund, 1999].

In conclusion, the paper proposes a procedure that makes CSI applicable to most real situations, where the required disease model parameters are not known. The CSI sets thus constructed are precise and have at least the nominal coverage level for reasonable sample sizes. In the application of the two-step procedure to a real data example, we see that the method provides sufficiently narrow confidence sets, which can be used as a starting point for further gene mapping studies.
Chapter 4

Alternative Confidence Sets for Disease Gene Locations

The LOD support interval approach is the oldest way of constructing confidence intervals for disease gene locations in linkage analysis. While it has sound statistical properties in single-point linkage analysis (for confidence intervals for $\theta$, the recombination fraction), its properties in multi-point linkage are less well understood [Dupuis and Siegmund, 1999, Papachristou and Lin, 2006c]. The method of estimating the disease gene location using Generalized Estimating Equations (GEE) proposed by Liang et al. [2001a] offers an alternative in linkage analysis of ASP data. The GEE method has been studied in some detail and extended to encompass a number of different scenarios. Bootstrap may be used to construct confidence intervals for parameters of interest including the location of a disease gene [Papachristou and Lin, 2006c]. In addition to CSI, GEE, Bootstrap methods, and LOD support intervals, all widely used or more fully developed approaches, there exist other less studied methods (e.g., Hauser et al. [1996], Hössjer [2003]).

In this chapter, we will review the LOD support interval method and the GEE method in some detail. Some shortcomings of the GEE method are studied and a comparison of these
methods for linkage analysis of ASP data is reviewed.

4.1 LOD Support Intervals

Geneticists have traditionally used the LOD-drop approach to construct confidence intervals for the recombination fraction, $\theta$, between the putative disease locus and a marker, and consequently for disease gene location. In two-point linkage analysis in model-based linkage analysis, the LOD-drop approach is motivated as follows. Let $L(\theta)$ be the likelihood for the observed phenotype and genotype data of $\theta$. Then, the null hypothesis, $H_0 : \theta = 1/2$, may be tested using the likelihood ratio test statistic $-2 \ln(\Lambda) = 2 \ln[L(\hat{\theta})/L(1/2)]$, $\hat{\theta}$ denoting the maximum likelihood estimator (MLE) of $\theta$. Under the null hypothesis, $-2 \ln(\Lambda)$ follows a 50:50 mixture of two $\chi^2$ distributions with 0 and 1 degrees of freedom. The reason for the mixture distribution is that the null value, $\theta = 1/2$, lies on the boundary of the parameter space and not in its interior [Self and Liang, 1987]. The relation between the LOD-score and the likelihood ratio test is given by $LOD(\theta) = -2 \ln(\Lambda)/(2 \ln(10))$.

Consider the alternative interpretation of a confidence interval. The $(1 - \alpha)$ confidence interval for $\theta$ contains all those $\theta_0$ such that the null hypothesis, $H_0 : \theta = \theta_0$, could not be rejected at the $\alpha$ level. That is, the 95% confidence interval for $\theta$ contains all $\theta_0$ that satisfy the inequality

$$2 \ln \left( \frac{L(\hat{\theta})}{L(\theta_0)} \right) < \chi^2_{1,0.05} = 3.84$$

$$LOD(\hat{\theta}) - LOD(\theta_0) < \frac{3.84}{4.6} = 0.835.$$ 

Hence, we see that the set $S = \{\theta_0 : LOD(\theta_0) > LOD(\hat{\theta}) - 0.835\}$ is a 95% confidence set for $\theta$. Traditionally, geneticists have used a drop of 1, as opposed to 0.835, and this corresponds to a 96.8% confidence set. Ott [1999] warns that the above confidence set is
valid only when $LOD(\hat{\theta}) > 3$. Figure 4.1 shows how a 95% confidence interval for $\theta$ may be constructed. Once a confidence interval for the recombination fraction between a known marker and the putative disease locus is obtained, it can be converted to a corresponding region (depending on distance from known marker) on the chromosome.

For genomewide analysis support intervals provide a method of estimating the location of a trait locus [Dupuis and Siegmund, 1999]. Let $LOD(l)$ now denote the LOD score at the location $l$ on the genome. Given a value $d > 0$, a support interval is given by

$$\{ q : LOD(q) > \max_l LOD(l) - d \}.$$

As mentioned earlier, a $d$ of 1 is often used. While for a single marker a 1-LOD support interval is a 96.8% confidence interval, this results does not generalize to genome-wide scans involving reasonably dense markers. In such a situation the coverage of the above support interval will depend on the density of markers and the strength of the signal at

![Figure 4.1: The LOD-score curve and the corresponding LOD-drop confidence interval. A drop of 0.835 leads to a 95% confidence interval, denoted by the gray line.](image)
the disease locus. Dupuis and Siegmund [1999] show through theoretical analysis and a simulation study that for a quantitative trait locus, the 1-LOD drop and the 1.5-LOD drop correspond to a 90% and a 95% confidence region, respectively (for a dense map of markers, one every 1 cM). So, while a support region is approximately a confidence region, its confidence is different from that suggested by standard statistical theory.

4.2 Disease Gene Localization with Generalized Estimating Equations

Liang et al. [2001a] proposed a novel method of estimating the location of an underlying disease gene using Generalized Estimating Equations (GEE) [Liang and Zeger, 1986]. Traditionally, linkage analysis has been seen as a hypothesis testing problem, with the hypothesis of “no linkage” being tested at each marker or location on the genome. The multi-point model-free linkage analysis approaches to testing this hypothesis creates the temptation to conclude that the map location giving the maximum evidence against the null hypothesis represents the most likely location for the susceptibility locus. However, the magnitude of this evidence depends on, among other things, the sample size and the informativeness of the individual genetic marker. Liang et al. [2001a] proposed a method to estimate the location of an unobserved susceptibility gene when there is preliminary evidence that the chromosomal region framed by multiple markers includes a single disease gene. The method is based on the IBD statistic, that is, the number of alleles shared IBD, and hence does not need specification of mode of inheritance. The method is also geared to use information from multiple markers.
4.2.1 The Approach of Liang et al. [2001a]

Let $R$ denote the chromosomal region of length $T$ cM to which there is evidence of linkage. In the region there are $M$ typed markers at the loci $0 \leq t_1 \leq \cdots \leq t_M \leq T$. The method tries to infer linkage based on allele sharing among affected siblings. Let $n$ denote the number of pedigrees enrolled, and $K_i$ be the number of pairs of affected siblings in the $i^{th}$ pedigree ($i = 1, \cdots, n$). Let $S_{ik}(t)$ denote the number of alleles (0, 1 or 2) shared IBD at a location $t$, $0 \leq t \leq T$ for the $k^{th}$ ASP in the $i^{th}$ pedigree. Since the markers may not be fully informative, the IBD sharing must be deduced from the marker data $Y_i = \{Y_i(t_1), \cdots, Y_i(t_M)\}$ for the $i^{th}$ pedigree. Here, $Y_i(t_j)$ denotes the available marker data for the $i^{th}$ pedigree at locus $t_j$, ($j = 1, \cdots, M$). This information combined with population allele frequencies and the method of Kruglyak et al. [1996], can be used to compute $P(S_{ik}(t) = l|Y_i)$ and $S^*_i(t) = \sum_{l=0}^{2} lP(S_{ik}(t) = l|Y_i)$, the conditional expectation of the number of alleles shared IBD at the location $t$ ($k = 1, \cdots, K_i$ and $i = 1, \cdots, n$).

If there is only one susceptibility locus in $R$, Liang et al. [2001a] showed that, under the assumptions of random mating, linkage equilibrium, and generalized single ascertainment, the expected number of alleles shared IBD at a fully informative marker locus $t$ has the form

$$\mu(t) = E(S^*(t)|B) = 1 + (2\Psi - 1)[E(S(\tau)|B) - 1],$$

(4.1)

where $B$ denotes the event that both siblings are affected, and the term $\Psi = \theta^2 + (1 - \theta)^2$, with $\theta$ being the recombination fraction between marker $t$ and the true disease location $\tau$. Then formula (4.1) represents the IBD sharing at $t$ as a function of recombination between $t$ and $\tau$ and expected alleles sharing at $\tau$ through $C = E(S(\tau)|\Phi) - 1$.

The first element of $\mu(t)$ involves recombination and Liang et al. [2001a] used the familiar
map function of Haldane [1919] to rewrite (4.1) as

\[ \mu(t; \tau, C) = 1 + e^{-0.04|t-\tau|}C. \]  

(4.2)

The second term involves \( C \), which lies between +1 and -1. An interpretation of \( C \) is that \( (1 + C)/2 \) characterizes the probability that an affected sibpair shares the same allele at \( \tau \) from the parent. A \( C \) value of zero means that there is no linkage of region \( R \) to the disease locus. The value of \( C \) depends on the genetic model but it is not possible to identify the genetic model from \( C \). The greater the value of \( C \) the easier it is to estimate \( \tau \).

To estimate \( \delta = (\tau, C) \), Liang et al. [2001a] solved the following estimating equations,

\[ \sum_{i=1}^{n} \sum_{k=1}^{K_i} \left( \frac{\partial \mu(\delta)}{\partial \delta} \right)' \text{Cov}^{-1}(S_{ik}(Y_i)|\Phi_i) (S_{ik}(Y_i) - \mu(\delta)) = 0, \]  

(4.3)

where \( S_{ik}(Y_i) = (S^*_{ik}(t_1), \ldots, S^*_{ik}(t_M))' \) and \( \mu(\delta) = (\mu(t_1; \delta), \ldots, \mu(t_M; \delta))' \). The theory developed by Liang and Zeger [1986] ensures that as long as the mean of the observations is correctly specified (i.e., \( \mu(t) = E(S^*(t)|B) \)) the GEE estimators of \( \tau \) and \( C \) converge towards the true location position and genetic effects as the sample size increases.

A specification of \( \text{Cov}^{-1}(S_{ik}(Y_i)|\Phi_i) \) in terms of the unknown \( \tau \) and \( C \) and the assumption of completely informative markers is given by Liang et al. [2001a], but is not essential for the consistency of the procedure; it only affects its efficiency. In addition, an asymptotically robust variance-covariance matrix for the estimates \((\hat{\tau}, \hat{C})\) can be computed [Lebrec et al., 2006]. The standard error of the estimate can be used to construct asymptotic confidence intervals for the disease location.

The proposed method is applied to simulated data sets with varying number of pedigrees (50, 100, 200 ASPs) and different map densities (microsatellite markers every 10 or 20 cM). Increasing sample size led to narrower confidence intervals for the disease location \( \tau \) and \( C \). In conclusion, the paper provides a novel method for localizing underlying disease
genes, after preliminary evidence of linkage has been obtained in a region, using the GEE methodology. The method incorporates information from multiple markers and, in addition to Hardy-Weinburg equilibrium and linkage equilibrium between markers, it only assumes that there is a single disease locus in the region of study.

### 4.2.2 Extensions

The method of estimation of disease gene location using the GEE methodology [Liang et al., 2001a] has received considerable attention. There have been a number of papers devoted to extending the method to more general situations. Liang et al. [2001b] extended the method to assess evidence of linkage to one region by incorporating linkage evidence from an unlinked region. In complex diseases, covariate data may be available that reflect etiologic or locus heterogeneity. Glidden et al. [2003] and Chiou et al. [2005] incorporate covariate information into the method by writing $C$ as a function of the observed covariate, and estimating the functional relationship using the estimating equations. Biernacka et al. [2005] extend the GEE method to simultaneously localize two linked disease susceptibility genes. Schaid et al. [2005] extend the method to different types of affected relative pairs.

### 4.2.3 Coverage of Confidence Intervals

The constructed confidence intervals for the disease gene location have been observed to have less than nominal coverage [Lebrec et al., 2006, Papachristou and Lin, 2006c]. As Lebrec et al. [2006] points out, the GEE methodology depends on the correct specification of the mean ($\mu(t) = E(S^*(t)|B)$); however, the proof of Liang et al. [2001a] assumes that the marker $t$ is completely informative and this assumption is rarely satisfied in real data. Lebrec et al. [2006] show with simulated data that, in regions where IBD information is sparse, in particular, where the multipoint information from markers varies widely.
the GEE method can lead to biased estimates of the location and hence less than nominal coverage of confidence intervals. This improper coverage was observed for both the asymptotic confidence intervals and bootstrap confidence sets. Lebrec et al. [2006] argue that highly informative dense SNP maps with less variable information content could increase the reliability of the GEE method. While the increase in information helps linkage signal localization, it still does not guarantee that the mean structure is correctly specified in the GEE model. Under-coverage of the GEE method with dense SNP maps has been observed by Papachristou and Lin [2006b] and our own results.

4.3 Comparison of Intermediate Mapping Methods

Whole genome studies are usually followed by association fine mapping methods focusing on the regions that house linkage signals. This two stage protocol would be more efficient if, between them, one constructed precise confidence sets with correct coverage for the disease gene location using dense SNPs from the linkage data. Papachristou and Lin [2006c] dub this the “intermediate fine mapping” step. In addition to having to follow up a smaller region with association mapping, it lends objectivity to the whole process. This revised protocol also helps the multiplicity adjustment issue in the later stages of the analysis. As more and more dense, yet cost-effective, SNP marker maps become available, we can shift the focus of linkage analysis from testing for linkage signals to sufficiently localize positions of putative disease loci. Three methods, already reviewed in detail, which provide confidence sets for the underlying disease gene location, are the CSI procedure, the GEE method and the LOD-drop approach. Papachristou and Lin [2006c] compare the performance of the above three methods in addition to two proposed bootstrap methods for constructing confidence intervals for the disease gene location. The only assumption common to all methods is that the preliminary broad linkage region (PBLR) houses at most one disease contributing gene.
4.3.1 Methods

As mentioned earlier, five methods have been compared in the paper: they are the LOD-drop approach, the GEE method, the CSI approach and two proposed bootstrap methods. Here are details of the methods used.

**LOD support interval:** The authors used the equivalent LOD score of Kong and Cox [1997] to construct LOD support intervals with a drop size of 1. The drop size of 1 was chosen because it has been traditionally used to provide 95% confidence intervals for the disease gene location.

**CSI approach:** The two-step CSI approach using the Mean test [Papachristou and Lin, 2006b] may be used to construct confidence sets for the disease gene location. This version of the CSI does not require knowledge of disease parameters and hence is easily applicable to real data.

**GEE method:** Also compared is the method of Liang et al. [2001a] which uses generalized estimating equations to estimate the location of the disease causing gene.

**Bootstrap approaches:** Two bootstrap approaches based on the likelihood method proposed by Risch [1990c] are proposed. The maximization, when the IBD sharing probabilities at the disease locus are contrained to the possible triangle [Holmans, 1993] leads to a consistent estimator of the true location of the putative gene, \( \tau \), [Risch, 1990c, Holmans, 1993, Cordell, 2004]. However, the constrained maximization means that the MLE, \( \hat{\tau} \), may no longer be asymptotically normally distributed. To overcome this problem, Papachristou and Lin [2006c] proposed two bootstrap methods to construct the empirical distribution of the MLE of the location. The first of the two methods is the *Non-Parametric Bootstrap*
(NPB), and involves sampling $n$ ASPs with replacement from the original pairs. Based on this bootstrap sample, the MLE of the location $\hat{\tau}_b$ is obtained. The procedure is repeated $B$ times to obtain a sample of bootstrap estimates $\{\hat{\tau}_b, b = 1, \ldots, B\}$, which may be regarded as an empirical distribution of $\hat{\tau}$. The $\alpha/2$ and the $(1 - \alpha/2)$ quantiles of this distribution provide a $(1 - \alpha)$ confidence interval for $\tau$. An alternative procedure proposed is the Parametric Bootstrap (PMB). Here instead of resampling ASPs with replacement from the original set, the method of Papachristou and Lin [2006a] is used to simulate pedigrees with the same structure and disease model as the MLEs of the location and IBD sharing probabilities of the original sample.

4.3.2 Simulation and Results

The four two locus disease models in Papachristou and Lin [2006b] are simulated with two disease chromosomes of lengths 150 and 100 cM. The true disease locus is close to the middle of each chromosome, at 75.2 and 52.7 cM, respectively. Both chromosomes had MS markers (every 7.5 cM) and SNP markers (every 0.5 cM). Samples with 100, 250, and 500 independent ASPs were analyzed. For each chromosome the family genotypes on the MS markers only were extracted, and used to perform a scan using the KAC statistic [Kong and Cox, 1997], as implemented in Allegro [Gudbjartsson et al., 2000], to uncover evidence of linkage. The threshold for declaring linkage was set to be 2.33, since this cut-off is very often used in preliminary linkage studies to flag suggestive signals [Kruglyak et al., 1995]. Replicates that yielded a maximum KAC-score above the chosen threshold for at least one of the two chromosomes were retained, with the PBLR taken to be composed of all positions within 25 cM from the location of the maximum KAC score. Five hundred replicates of each type are analyzed. In the second step, the intermediate fine mapping stage, attention is focused on the PBLR identified from the preliminary scans. Specifically, using all SNP markers within the PBLR, confidence intervals using each of the five methods
is constructed.

The coverage probability is defined as the ratio of the number of intervals that captured the location of the trait locus, over the number of replicates that signaled linkage on the chromosome. The two bootstrap methods provided confidence regions with coverage probabilities close to the nominal 95%. For small sample sizes, PMB was slightly liberal, leading to reduced coverage. The two-step CSI procedure had greater than nominal coverage, reaching almost 100% for all models even for small sample sizes. This is partially due to the fact that confidence regions were constructed only for those replicates that passed the initial screening for linkage. In doing so, it is likely that the chance of the interval including the trait locus increased, resulting in a marginally elevated coverage. But this increased coverage is also because of the correlation in the IBD sharing information provided by the MS and the SNP map. Thus, the CSI procedure led to wider intervals with higher coverage probability.

In contrast to CSI, LOD support intervals are known to provide short intervals with lower than nominal coverage [Benewitz et al., 2002]. This is especially true when there is a limited amount of data. The results of the coverage of the LOD support intervals agree with these previous observations. However, it is surprising to see that in some situations, this under-coverage is very adverse, for example, in one situation the 1-LOD drop has a coverage of only 26%. The GEE method is also known to produce confidence intervals that significantly undercover the disease gene location [Biernacka et al., 2005, Lebrec et al., 2006]. This under-coverage was also observed in the results of the paper, though it is less severe than that of the 1-LOD drop interval. This undercoverage was more pronounced for smaller sample sizes. However, even with 500 ASPs, they rarely exceeded 90%.

The huge variation in the coverage of the different methods means that they cannot be compared in the sense of the precision of the confidence intervals. So, the different methods were contrasted after maintaining the observed coverage to the correct level for all methods.
The results suggest that, when the samples were small to moderate (100 to 250 ASPs), and the true coverage probabilities were held to the same level for all methods, CSI produced confidence regions that tended to be shorter than those obtained through any of the other four methods. However, as the sample size increased to 500, LOD-drop took the lead as the method better localizing the trait loci than any of the other methods.

For the two bootstrap methods, the NPB produced intervals that typically were wider than those obtained by any of the competing methods. This is to be expected, since, as a pure non-parametric approach, it lacks power to significantly localize the trait loci. However, the PMB gave very encouraging results when the number of ASPs was relatively high with 500 pairs. For almost all the trait loci, it provided confidence intervals that, on average, were only slightly wider (about 2cM) than those derived by the best method.

Finally, the GEE method demonstrated a very low capability to localize trait loci. It consistently yielded very long intervals, with average lengths quite often exceeding those obtained by the NPB. In particular, it experienced difficulties in isolating trait loci with small contribution to the trait, especially when the amount of data was as small as 100 ASPs. However, even with 500 ASPs, the method rarely outperformed the NPB for the task of intermediate fine mapping.

In conclusion, the bootstrap methods tend to provide intervals with correct coverage. LOD-drop and GEE methods provide under-coverage of the disease gene location. The CSI procedure on the other hand provides wide intervals with greater than nominal coverage. When constrained to the same observed coverage, the CSI method performs the best for small sample sizes, while the LOD method provides the shortest confidence intervals when the sample size is 500 ASPs.
Chapter 5

Bayesian Methods in Linkage

Bayesian methods in statistics provide a general framework in which many of the problems in genetics can be studied. The sheer amount of computational power that is required for most Bayesian statistical analyses has previously limited their use in genetics. The vast increase in easily available cost effective computational resources means that these computational constraints have now largely been overcome. Bayesian statistics provide scientists with two primary advantages. The first is that Bayesian analysis provides the ability to incorporate background information into the specification of the model, in the form of a prior. However, as Beaumont and Ranala [2004] argue, the recent popularity of Bayesian methods can be explained largely by the relative ease with which complex likelihood problems can be tackled by the use of computationally intensive Markov Chain Monte Carlo (MCMC) techniques. These underlying advantages of Bayesian approaches have brought them to the forefront of genetic data analysis.

In this Chapter, we will review the principles of Bayesian inference, followed by a look at some of the applications of Bayesian methods to linkage analysis. Finally, we will see some examples to Bayesian credible intervals for disease gene locations.
5.1 Principles of Bayesian Statistics

The essence of the Bayesian viewpoint is that there is no philosophical distinction between data and the model parameters. Both are random variables with a joint probability distribution that is specified by a probabilistic model. The joint distribution is a product of the likelihood and the prior. The prior contains information about the possible values of a parameter, before the data is examined, in the form of a probability distribution. The likelihood is a conditional distribution that specifies the probability of the observed data given particular values of the parameters and is based on a model of the underlying process. Together, these two functions combine all available information about the parameters. Bayesian statistics involves working with this joint distribution in various ways to make inferences about the parameters, or the probability model, given the data. The main aim of Bayesian inference is to calculate the posterior distribution of the parameters, which is the conditional distribution of parameters given the data. A point estimate of a parameter is obtained by considering some measure of central tendency of the posterior distribution (usually the mode or the mean). An interval estimates of a parameter can be obtained by considering a credible set of values (a set or interval that contains the true parameter with probability $1 - \alpha$, for which $\alpha$ is a pre-specified significance level such as 0.05).

There are many practical reasons to use Bayesian inference: if a probability model includes many interdependent variables that are constrained to a particular range of values (as is often the case in genetics), maximum likelihood inference requires that a constrained multidimensional maximization be carried out to find the combined set of parameter values that maximize the likelihood function. This is often a difficult numerical analysis problem. In addition, under the maximum-likelihood method, calculation of confidence intervals and statistical tests generally involve asymptotic approximations to normality, and for small samples such an assumption may not be satisfied. Approximations to asymptotic distributions are further complicated under constrained maximizations [Self and Liang, 1987].
the other hand, in Bayesian inference, inferences about parameter values on the basis of the posterior distribution usually require integration (for example, calculating means) rather than maximization, and no further approximations are involved.

MCMC methods rely on constructing a Markov chain with a stationary distribution that is the probability distribution of interest. Inference about the distribution can then be drawn by performing Monte Carlo integration on a (practically independent) sample from this target distribution. In Bayesian analysis, this target distribution is often the posterior distribution of the parameters. Modern MCMC methods implemented on powerful new computers have greatly facilitated the computation of Bayesian posterior probabilities, making the calculations tractable for complicated genetic models.

5.2 Bayesian Methods in Linkage Analysis

Use of Bayesian methods in linkage analysis has a long history. In his seminal work describing the LOD score method, Morton [Morton, 1955] was interested in the probability that a marker is truly linked when we conclude that a linkage is present. This probability may also be called 1 minus the posterior type I error. Let ‘+’ denote the presence of true linkage and ‘-’ denote non-linkage to a particular region. Let $L$ denote the rejection of the null hypothesis (the conclusion that there is linkage). Then, the type I error is $P(L|-) = \alpha$ and the power is $P(L|+) = 1 - \beta$. Under the assumption that there is a trait locus, Morton set a value of $P(+) = 0.05$, that is, the prior probability that any two loci will be on the same chromosome is $1/22 \approx 0.05$. Then the probability that a linkage is correct when we declare linkage is

$$P(+|L) = \frac{(1 - \beta)P(+)}{(1 - \beta)P(+) + \alpha[1 - P(+)].}$$

(5.1)

For the derivation please see Rao and Province [2000]. It is easy to see that, with $\alpha = 0.05$ and $\beta = 0$, $P(+|L) = 0.51$. That is, even with 100% power, only about half the reported
linkages would be real. This led Morton to choose $\alpha = 0.001$ to reduce the posterior type I error.

Vieland [1998] concentrates on another quantity, the Posterior Probability of Linkage, which may be defined as $\text{PPL} = P(+|D)$, where $D$ is the observed genotype and phenotype data. The PPL is given by the integral

$$\text{PPL} = \int_{0\leq \theta < 1/2} f(\theta|D) d\theta, \tag{5.2}$$

where, $\theta$ is the recombination fraction between the marker locus and the trait locus and $f(\theta|D)$ is the posterior distribution of $\theta$. The PPL is a quantity of considerable interest to a geneticist at the end of the study [Smith, 1959, Elston, 1997], and as Vieland [1998] showed, it can be obtained with relative ease. The PPL also has the advantage that new data can incorporated into the method to obtain an updated PPL. Through the use of Bayesian priors the PPL allows for an unknown genetic model without fixing it a priori and without maximization. In effect, the PPL determines the average amount of evidence across the entire genetic model space, with the weight given to different portions of the parameter space being controlled by the priors. However, the PPL does not treat the location of the underlying disease gene as a parameter in the model and does not provide inference about it.

The computational burden in likelihood methods is considerably larger for model-based linkage analysis than for segregation analysis and grows rapidly with the number of loci, number of alleles, and pedigree size and complexity (especially if there are untyped individuals) Thomas [2004]. Even with a single locus, joint segregation and linkage is seldom possible. MCMC methods offer potentially attractive solutions. Heath [1997] created the program LOKI, which conducts multi-point linkage analysis for a quantitative trait in general pedigrees allowing for an unknown number of trait loci. At each iteration, the location of each locus are allowed to move around, their segregation parameters are updated, and
new loci can be added to the model or existing loci removed, using the technique of reversible jump MCMC [Green, 1995]. The output of the chain may be sampled to construct the posterior distribution of the parameters of interest including the locations of trait causing gene. A credible interval can hence be constructed. The credible interval may be seen as a starting point to perform further fine mapping.

In spirit, the Bayesian joint segregation and linkage analysis for quantitative traits [Heath, 1997] can be extended to binary traits. However, the primary reason for not estimating all parameters jointly for qualitative trait linkage is that the heavily loaded pedigrees that are the most informative for linkage are seldom a representative sample of all families. Therefore, any attempt to estimate the segregation parameters from such highly ascertained families is likely to lead to biased estimates. For these reasons, these parameters are often fixed at values determined by segregation analyses of other, more representative, families.

5.3 Bayesian Credible Intervals for Disease Gene Locations

5.3.1 Linkage Disequilibrium Based Fine Mapping

Alleles at loci that are in close proximity on a chromosome may exhibit a non-random association known as Linkage disequilibrium (LD). LD mapping is based on the idea that a disease mutation arises on a particular chromosome and is initially associated with marker alleles on that chromosome. Over time, recombination breaks this association down, narrowing it only to markers that are very close to the disease mutation on the chromosome. To map causal mutations using LD, a population sample of haplotypes, or genotypes, from unrelated individuals that are either affected, or unaffected, by a disease is collected. Often, a set of markers are chosen that span a larger candidate region identified first by conven-
tional family-based linkage analysis. While linkage analysis can successfully localize a disease locus to within some centi Morgan, LD mapping can presumably narrow the region down to hundreds of Kilo Bases. Construction of confidence intervals for disease genes using LD mapping methods, at a resolution greater than that possible with linkage analysis, has received considerable attention in the recent years. Bayesian [Morris et al., 2002, Molitor et al., 2003, Waldron et al., 2006] and frequentist [Zöllner and Prichard, 2005] methods have been proposed for constructing credible/confidence regions for the location of a disease gene with LD mapping.

5.3.2 Quantitative Trait Locus Mapping

In Quantitative Trait Locus (QTL) mapping in experimental organisms, Dupuis and Siegmund [1999], compared three ways of constructing confidence/credible regions for the location of the underlying QTL. They compared the LOD support intervals, confidence regions from the likelihood and Bayesian credible intervals. The authors argue that in the problem of estimation of QTL location, a Bayesian credible region for a uniform prior distribution on the location will provide satisfactory confidence regions. In the particular situation considered, the posterior distribution can be computed in a closed form. Three different priors are considered for the parameters denoting the effect size of the QTL. A comparison of the three ways for constructing 95% confidence or credible regions shows that, the LOD support interval (with a drop of 1.5) gives the smallest confidence region for the dense map, while the Bayesian credible interval did the same for the sparse map.
Part II

Two Novel Approaches To Construct

Efficient Confidence Sets
In positional cloning of disease causing genes, identification of a linked chromosomal region via linkage studies is often followed by fine mapping via association studies. Efficiency can be gained with an intermediate step where confidence regions for the locations of disease genes are constructed. The Confidence Set Inference (CSI) [Papachristou and Lin, 2006a] achieves this goal by replacing the traditional null hypotheses of no linkage with a new set of null hypotheses where the chromosomal position under consideration is in tight linkage with a trait locus. This approach was shown to perform favorably compared to several competing methods [Papachristou and Lin, 2006c]. Using the duality of confidence sets and hypothesis testing, CSI was proposed for the Mean test statistics with affected sibling pair data (CSI-Mean). We postulate that more efficient confidence sets will result if more efficient test statistics are used in the CSI framework. One promising candidate, the Maximum LOD Score (MLS) statistic, makes maximum use of available identity by descent information, in addition to handling markers with incomplete polymorphism naturally. We propose a procedure that tests the CSI null hypotheses using the MLS statistic (CSI-MLS). Compared to CSI-Mean, CSI-MLS provides tighter confidence regions over
a range of single and two-locus disease models. The MLS test is also shown to be more powerful than the Mean test in testing the CSI null over a wide range of disease models, the advantage being most pronounced for recessive models. In addition, CSI-MLS eliminates the need for Monte Carlo simulations needed in CSI-Mean, and hence is computationally more efficient.

The CSI framework requires knowledge of some disease model related parameters. These parameters can possibly be estimated from epidemiologic studies. In this Chapter we shall assume that the required parameters are known, and compare CSI-MLS and CSI-Mean on the effect of a number of interesting factors to gain a comprehensive understanding. In the following Chapter, we implement CSI-MLS in the two-step CSI procedure suggested by Papachristou and Lin [2006b].

### 6.1 More Efficient CSI with Likelihood Ratio Tests: CSI-MLS

As reviewed in Chapter 3, CSI is a general framework whose hypotheses can be tested by reformulation of any reasonable traditional linkage test statistic. CSI-Mean is a logical first step given the popularity and early application of the Mean test. In view of CSI’s nature of obtaining the confidence sets via inverting the acceptance regions of hypothesis tests, we postulate that tests that are more powerful in traditional linkage studies might carry this advantage over to testing the CSI nulls and thus result in more efficient confidence set estimators. The MLS [Risch, 1990c] was one of the first model free linkage methods that can properly account for the uncertainty in the number of alleles shared IBD. Furthermore, it makes maximum use of the information available from allelic identity by descent. Davis and Weeks [1997] compared a number of linkage test statistics and showed that eight of the ten best statistics were likelihood based tests (the other two being the Mean test and the
Haseman-Elston regression). We derive here the CSI-MLS that tests the CSI nulls using the MLS, and discuss some of the immediate benefits before we compare its efficiency to that of CSI-Mean.

6.1.1 Single-point CSI-MLS

Recall that the hypothesis for each marker $m$, in single-point CSI, is

\[ H_{0m}^{(\text{new})}: \theta_m \leq \theta_0 \text{ versus } H_{Am}^{(\text{new})}: \theta_m > \theta_0. \]  

(6.1)

Let $mIBD$ be the number of alleles shared IBD by an affected sibling pair (ASP) at marker $m$. Let $z_i(\theta) = P(mIBD = i | B, \theta)$, where $\theta$ is the distance between the trait locus and marker $m$. Further, let $G_j^m$ be the genotype data at marker $m$ for $j^{th}$ ASP (or family). The likelihood is [Risch, 1990c]

\[
L(z_0(\theta), z_1(\theta), z_2(\theta)) = \prod_{j=1}^{n} P(G_j^m | z_0(\theta), z_1(\theta), z_2(\theta)) \\
= \prod_{j=1}^{n} \sum_{i=0}^{2} P(G_j^m | mIBD_j = i) z_i(\theta) \\
= \prod_{j=1}^{n} \sum_{i=0}^{2} \frac{P(mIBD_j = i | G_j^m) P(G_j^m)}{P(mIBD_j = i)} z_i(\theta)
\]

(6.2)

Define $w_{ij} = P(mIBD_j = i | G_j^m) P(G_j^m) / P(mIBD_j = i)$. Then the likelihood ratio test (LRT) statistic for hypothesis (6.1) is

\[
\Lambda_m = \frac{\sup_{\theta \leq \theta_0} \prod_{i=1}^{n} (\sum_{i=0}^{2} z_i(\theta) w_{ij})}{\sup_{\theta \leq 1/2} \prod_{j=1}^{n} (\sum_{i=0}^{2} z_i(\theta) w_{ij})}.
\]

(6.3)
For a single-locus disease model, the probabilities $z_i(\theta)$ can be expressed in terms of $\lambda_O$, $\lambda_S$ and $\theta$ [Risch, 1990b],

$$
z_0(\theta) = \frac{2\Psi - 1 + 2(2\Psi - 1)(1 - \Psi)\lambda_O + 4(1 - \Psi)^2\lambda_S}{4\lambda_S},
$$

$$
z_1(\theta) = \frac{2(2\Psi - 1)^2\lambda_O + 8\Psi(1 - \Psi)\lambda_S}{4\lambda_S},
$$

$$
z_2(\theta) = \frac{1 - 2\Psi - 2\Psi(2\Psi - 1)\lambda_O + 4\Psi^2\lambda_S}{4\lambda_S},
$$

(6.4)

where $\Psi = \theta^2 + (1 - \theta)^2$. The parameters $\lambda_O$ and $\lambda_S$ can either be estimated from independent genetic epidemiological studies or from an initial coarse linkage scan. By representing $z_i(\theta)$ as above, we have written likelihood (6.2) in terms of a single unknown parameter $\theta$. The maximization in the numerator and denominator of (6.3) is linear in the number of ASPs and thus computationally fast. The parametric estimates of $z_i(\theta)$ obtained in this manner satisfy the possible triangle constraints [Holmans, 1993]. A proof is available in the appendix (B.0.8).

Of the probabilities involved in (6.2), $P(G_{m}^{\text{null}})$ cancels in the LRT (6.3). The probabilities $P(m|IBD_j = i)$ are simply the traditional null IBD sharing probabilities, for siblings they are (1/4, 1/2, 1/4) for sharing (0, 1, 2) alleles IBD, respectively. The third probability $P(m|IBD_j = i|G_{j}^{\text{null}})$ depends on the observed genotypes of the $j^{th}$ family. These probabilities are well known quantities [Haseman and Elston, 1972] and are provided by a number of genetic software (e.g., S.A.G.E. [2006], Merlin [Abecasis et al., 2002], and Allegro [Gudbjartsson et al., 2000]).

Standard regularity conditions are not satisfied for the LRT (6.3) to have a simple chi-squared distribution. The asymptotic distribution of the statistic $-2 \ln \Lambda_m$ is a 50:50 mixture of $\chi^2_0$ and $\chi^2_1$ (our situation satisfies the conditions of Case 5 of Self and Liang [1987]). Assuming $\theta_0$ to be half the maximum distance between consecutive markers, the set $A = \{m : H_{0m}^{\text{new}} \text{ is not rejected at level } \alpha\}$ contains at least one marker within distance $\theta_0$ of the
disease locus with probability $\geq 1 - \alpha$. The confidence region $R = \cup_{m \in A} [\tau_m - \theta_0, \tau_m + \theta_0]$, \(\tau_m\) being the location of marker \(m\), contains the disease locus with probability $\geq 1 - \alpha$ [Papachristou and Lin, 2005a].

A distinction between the single-point CSI-MLS and CSI-Mean is their requirement for stochastic ordering in their respective test statistic. While there is no such requirement for CSI-MLS, the test statistic in the single-point CSI-Mean needs to be stochastically decreasing in $\theta$ in order for the test to have the correct size. Such a property is intuitive and was treated heuristically in Papachristou and Lin [2005a]. A mathematical proof of the stochastic ordering for completely informative markers is provided in the last Chapter.

### 6.1.2 Multi-point CSI-MLS

Genotypes at all markers on a chromosome potentially provide additional information about IBD sharing at each location on a chromosome. Such information can be extracted using a Hidden Markov Model [Lander and Green, 1987]. For a Microsatellite map, this means that the CSI test can now be conducted at locations between markers. With the advent of high throughput SNP technology we now have maps with thousands of low polymorphism markers. While individually not very informative, when information is extracted from them in a multi-point fashion these maps can be extremely useful for detecting linkage. Hence, the null hypothesis in multi-point CSI is, at each location ($\tau$) on the genome, that $\tau$ is the trait locus. Let $\theta_\tau$ be the recombination fraction between location $\tau$ and the disease locus $\tau^*$. Then the multi-point hypothesis is

$$H_{0_{\tau}}^{(\text{new})} : \theta_\tau = 0 \text{ versus } H_{A_{\tau}}^{(\text{new})} : \theta_\tau > 0,$$  \hspace{1cm} (6.5)
for every $\tau$ on the genome. The LRT statistic for hypothesis (6.5) is
\[
\Lambda_\tau = \frac{\prod_{j=1}^{n} (\sum_{i=0}^{2} z_i(0)w_{ij})}{\sup_{\theta \leq 1/2} \prod_{j=1}^{n} (\sum_{i=0}^{2} z_i(\theta)w_{ij})}.
\] (6.6)

The probabilities $z_i(\theta)$ are given in (6.4). Define $w_{ij} = P(\tau IBD_j = i | G_j)P(G_j)/P(\tau IBD_j = i)$, where $G_j$ denotes multi-point marker genotypes of the $j$th family. The terms $P(IBD_j = i | G_j)$ in $w_{ij}$ are multi-point probabilities which can be calculated using the Hidden Markov Model and are available from standard software packages. The distribution of $-2 \ln \Lambda_\tau$ is also a 50:50 mixture of $\chi^2_0$ and $\chi^2_1$. The maximization required is still one-dimensional and hence fast.

**Construction of confidence regions** The confidence region for the trait locus is given by the set $A = \{\tau : H_0^{(\text{new})} \text{ is not rejected at level } \alpha\}$. Since it is infeasible to perform the test at every point on the genome, we follow Papachristou and Lin [2006a] in conducting the hypothesis tests at a grid of points and then constructing the confidence region by interpolating the test statistics. Let the set of points at which tests are conducted be $\tau_l, l = 1, \cdots, L$. Let $T$ be a random variable with the same distribution as $-2 \ln \Lambda_\tau$, under the null hypothesis that $\tau = \tau^*$. Define $T_\alpha$ to be a threshold such that $P(T > T_\alpha) = \alpha$. A small value of $-2 \ln \Lambda_\tau$ is evidence in support of the null hypothesis. Therefore, in constructing the confidence set, first put in it all closed intervals $[\tau_l, \tau_{l+1}]$ such that $\max(-2 \ln \Lambda_{\tau_l}, -2 \ln \Lambda_{\tau_{l+1}}) < T_\alpha$. Do not include in the confidence set any intervals $[\tau_l, \tau_{l+1}]$ for which $\min(-2 \ln \Lambda_{\tau_l}, -2 \ln \Lambda_{\tau_{l+1}}) \geq T_\alpha$. In the situation where neither of the above two conditions is satisfied, interpolate $-2 \ln \Lambda_\tau$ with a linear function of $\tau$. If $-2 \ln \Lambda_{\tau_l} > (<) T_\alpha$ and $-2 \ln \Lambda_{\tau_{l+1}} < (>) T_\alpha$, then include in the confidence region the closed interval $[\tau', \tau_{l+1}]$ $([\tau_l, \tau'])$, where $\tau'$ is the point at which the line joining $(\tau_l, -2 \ln \Lambda_{\tau_l})$ and $(\tau_{l+1}, -2 \ln \Lambda_{\tau_{l+1}})$ crosses $T_\alpha$. 

72
6.1.3 Powers of CSI-MLS and CSI-Mean

We will investigate the range of disease models where the MLS and the Mean test have differing powers for the CSI hypothesis and potentially different efficiencies in CSI. The power of the MLS statistic to test hypothesis (6.5) is given by

\[ P(-2 \ln \Lambda > T_\alpha | \theta > 0) \]

A closed form of the power function of the MLS test is not available. We simplified the problem by considering a location \( \tau \) with complete information about IBD sharing and that is \( \theta \), away from the disease locus in a single-locus disease model. In this case, the number of pairs sharing (0, 1, 2) alleles IBD, \((n_0, n_1, n_2)\), out of \(n\) ASPs is distributed as a Multinomial\((n, z_0(\theta), z_1(\theta), z_2(\theta))\). Using this fact, Monte Carlo simulations can be performed to estimate the power of using MLS test statistic to test hypothesis (6.5).

While the asymptotic power function for a completely informative marker is available for the Mean test statistic [Lin, 2002], we obtained the Monte Carlo estimate of the power of CSI-Mean for consistency. Papachristou and Lin [2006a] suggested using a two-sided test for hypothesis (6.5). However, for a completely informative marker, a one-sided Mean test is appropriate and more powerful [Liang et al., 2001a, Papachristou and Lin, 2006a]. Hence, we used a one-sided CSI-Mean test when carrying out the comparison.

For a single-locus disease model, the power of both tests depend on three parameters, \((\theta, V_A/K^2, V_D/K^2)\), where \(K\) is the population prevalence, \(V_A\) is the additive genetic variance, and \(V_D\) is the dominant genetic variance. The power function should be increasing in all three parameters. Notice that \(V_A/K^2 = (1 - K)/K \times \) (Narrow Sense Heritability). Thus, the parameter \(V_A/K^2\) is a decreasing function of prevalence and an increasing function of heritability. A similar interpretation holds for \(V_D/K^2\). We can map \((V_A/K^2, V_D/K^2)\) into two parameters that are more directly related to the effect of disease models on power

\[
\eta = \frac{V_A/2 + V_D/4}{K^2 + V_A/2 + V_D/4} = \frac{\lambda_S \lambda_S - 1}{\lambda_S} \quad (6.7)
\]
\[
\zeta = \frac{V_D/4}{K^2 + V_A/2 + V_D/4} = \frac{\lambda_S - \lambda_O}{\lambda_S} \quad (6.8)
\]
The parameter $\eta$ lies between 0 and 1 and essentially contains information about the relative importance of the locus in the disease prevalence. The parameter $\zeta$ must lie between 0 and $\eta$ and contains information about how recessive the disease is [Feingold and Siegmund, 1997]. These parameters are closely related to the expected number of alleles shared IBD by an ASP at the disease locus, the relation being $z_1(0) + 2z_2(0) = (\eta + \zeta)/2$. The higher $\eta$ is, the more genetic the trait is, and the easier it is to reject the CSI null hypothesis at a marker that is not tightly linked to the disease locus.

### 6.1.4 Simulation Study Settings

Extensive simulation has been carried out to compare the efficiency of CSI-MLS and CSI-Mean, both single-point and multi-point. The effects of various factors have been studied. These factors include types of markers and the density of the marker map (Microsatellite and SNP maps), disease models with varying degrees of complexity and heritability, and sample sizes. One criterion of comparison is the length of the confidence set that contains the trait locus with a predetermined probability. In addition, we investigate the power of CSI-MLS and CSI-Mean tests as a function of $(\theta, V_A/K^2, V_D/K^2)$ and identify ranges of the disease models where one is preferred over the other. This power comparison is important in that the difference in power will translate into corresponding difference in efficiency of the confidence set estimators.

All the marker and trait loci considered are assumed to be in Hardy Weinberg Equilibrium (HWE) and in Linkage Equilibrium (LE) with each other. The trait locus is always assumed to be diallelic, and the two trait loci are assumed to be on different chromosomes when two-locus trait models are considered. The unit of simulation is an ASP with marker genotypes available for the ASP and their parents. The nominal coverage probability of the CSI confidence sets is set to be 95%. Unless otherwise noted, 1000 replicates under each setting are simulated.
In what follows, we briefly discuss the disease models employed in the simulation. These models have been studied extensively in Papachristou and Lin [2006a, 2006b]. Tables 6.1, 6.2, and 6.3 are adapted fully or partially from these two papers.

**Single-locus disease models** Three models with heritabilities ranging from high to low (71%, 50%, 14%), as displayed in Table 6.1, are used in the simulation to compare the efficiency of CSI-MLS and CSI-Mean, both single-point and multi-point. These models also cover the mode of inheritance ranging from recessive, additive, to dominant.

To compare the single-point CSI-MLS and CSI-Mean under these three models, we simulated 30 Microsatellite markers equally spaced on a 290-cM chromosome (10cM density). Each marker has ten equally frequent alleles. The disease locus is at 141 cM, 1 cM from the 15th marker. Each replicate comprises 250 nuclear families with two affected siblings.

We also compare the multi-point versions of the two procedures under these three single-locus trait models. In addition, the effect of having a Microsatellite (MS) map or a SNP map has also been investigated. The MS map comprises 17 MS markers equally spaced in a 120-cM chromosome (7.5 cM density), each with 8 equally frequent alleles. The trait locus is at 61.2 cM, 1.2 cM away from the nearest marker. The SNP map comprises one SNP every 0.25 cM along a 60-cM chromosome, each with a minor allele frequency of 0.3. The trait locus is at 30 cM. Sample sizes of 100, 250, 500, and 1000 (for Model III) are simulated.

**Two-locus disease models** Diseases of interest to genetic epidemiologists today are largely complex in nature. Four two-locus trait models (Tables 6.2 and 6.3), two epistatic and two heterogeneous, are considered. A fuller discussion of these models can be found in Papachristou and Lin [2006a, 2006b]. We simulate 17 MS markers equally spaced along a 120-cM chromosome for each of the two trait loci, each marker with 8 equally frequent
alleles. Each trait locus is at 61.2 cM on their respective chromosome. Sample sizes of 100 and 500 families are considered. For each setting we simulated 500 replicates.

6.2 Results

6.2.1 Comparison Between Single-point CSI-MLS and CSI-Mean

The CSI framework was developed to address the interest in better susceptibility gene localization in the era of dense marker map. Single-point CSI procedures are thus not of practical interest. However, they form the foundation of their multi-point counterparts and thus their investigation will aid the understanding of the more complex procedures. Figure 6.1 provides the relative frequency of each marker being included in the confidence sets constructed by CSI-MLS and by CSI-Mean. The mean length of the confidence sets in cM (M), the standard deviation of the lengths (SD), and the coverage probability (CP) are also provided. The mean length of the confidence sets are comparable for CSI-MLS and CSI-Mean. The probability of a marker being included in the confidence sets decreases as the marker gets farther away from the trait locus, the rate of decrease being faster for a model with higher heritability. A single-point confidence set does not distinguish between chromosomal positions that are symmetrically located on either side of a marker and this could lead to confidence regions wider than desired. This fact is reflected in the actual coverage probability of the confidence sets being larger than the nominal.

6.2.2 Comparison between multi-point CSI-MLS and CSI-Mean

We first discuss the simulation results under the three single-locus models given in Table 6.1.
**Effects of marker map, heritability, and sample size**

The results for a MS map of one Microsatellite marker per 7.5 cM is presented in Table 6.4. The results for a SNP map of one SNP per 0.25cM, designed to simulate commercially available chips (e.g., the 10K SNP arrays by Affymetrix or Illumina), are presented in Table 6.5. First of all, from Table 6.4 there is no systematic deviation in the coverage probability from the nominal level of 95% and hence the three methods can be compared in terms of precision, i.e., the length of the constructed confidence set. Similar consistency is observed for the results under the SNP map. For the MS map, two variants of CSI-Mean, CSI-Mean-V2 and CSI-Mean-V3, are compared to our proposed CSI-MLS. Unlike CSI-MLS or CSI-Mean-V3, CSI-Mean-V2 does not use IBD sharing probability estimates at locations between markers, reasoning that these may not be estimated precisely. This consideration seems to be advantageous for a highly heritable trait (Model I) under MS map, as reflected in CSI-Mean-V2 out-performing both CSI-MLS and CSI-Mean-V3 (Table 6.4). This advantage vanishes as we consider more realistic models with moderate heritability (Models II and III), where CSI-MLS consistently outperforms both variants of CSI-Mean by reducing the average length of the confidence sets by as much as around 20%.

For the SNP map (Table 6.5), CSI-Mean-V2 is not considered as there is no difference between CSI-Mean-V3 and CSI-Mean-V2 at this map density. CSI-MLS consistently out-performs CSI-Mean under all three models, reducing the average length of the confidence sets by between 7% to 20%.

Comparing results in Tables 6.4 and 6.5, it is apparent that the current panel of SNP maps provides confidence sets much more precise than those obtained from MS maps, sometimes reducing the length by more than 50%.

The underlying etiology of the disease also determines the precision with which one is able to localize the susceptibility gene. The more genetic a trait is, the easier it is to localize the susceptibility gene, as demonstrated in both Tables 6.4 and 6.5.
The reformulation of the tradition hypothesis to the CSI null that the location being tested is the trait locus guarantees that, asymptotically, a CSI test will produce an interval that will include only the trait location, provided the assumptions are correct. In other words, as the sample size increases, the length of a level \((1 - \alpha)\) confidence set decreases. This property of the CSI framework, using either CSI-MLS or CSI-Mean, is apparent in Tables 6.4 and 6.5. For each disease model and marker map, the length of the CSI intervals decreases as the sample size increases, irrespective of which version of CSI was used. Reviewing Table 6.5, by increasing the sample size from 100 to 250, the length of the confidence set is reduced to around 60%. A further increase of sample size to 500 can reduce the length to around 40%. For model I, 100 ASPs are more than enough to provide a tight region to be followed up on; whereas 250 ASPs are needed under model II and the sample size should be around 500 under model III.

**Effect of complexity of the disease model** To compare CSI-MLS and CSI-Mean in their efficiency in localizing the susceptibility genes involved in complex traits, the four two-locus models in Table 6.2 are simulated on a MS map with one marker per 7.5 cM. The results are presented in Table 6.6. As before, we assumed the exact IBD sharing at each of the trait loci by an ASP is known, which amounts to knowing the locus specific relative risks. Both variants of CSI-Mean were investigated and they yielded essentially the same result under each model. This is intuitive as, judging from the locus specific relative risks, none of the trait locus in the four two-locus models has stronger genetic effect than that of model II among the three one-locus models. Again, the coverage probabilities are approximately equal to the nominal level and thus the efficiency of the confidence set can be compared based on their lengths. CSI-MLS consistently provide confidence sets that are 11% to 23% shorter than that provided by CSI-Mean. The lengths are reduced to around 40% by increasing the sample size from 100 to 500.

The effect of the underlying disease model is clearly seen in the ability of the methods to
narrow trait locus regions. For the epistatic model EP-2, both loci have similar contribution and each method is able to localize the two loci with similar precision. For the other epistatic model EP-4, the 2\textsuperscript{nd} trait locus has a greater contribution to the disease and is identified with much greater precision. The lengths of the confidence sets are similar for the two loci in the heterogeneity model Het-2. For S-2, the 1\textsuperscript{st} trait locus has a much higher sibling relative risk than the 2\textsuperscript{nd} locus and is easier to localize. Comparing Tables 6.4 and 6.6, we observe that for each method, the confidence set length is approximately a decreasing function in \((\lambda_O, \lambda_S)\), as discussed before. For example, the effect of the 2\textsuperscript{nd} locus in Het-2 \((\lambda_O = 1.70, \lambda_S = 1.70)\) is slightly less than that of model III \((\lambda_O = 1.78, \lambda_S = 1.78)\), and this leads to a slightly wider confidence sets obtained from CSI-MLS (21.4 cM versus 20.8 cM with 500 ASPs).

### 6.2.3 Power of CSI-MLS and CSI-Mean

In the above two subsections, we have demonstrated that CSI-MLS is more efficient than CSI-Mean under three single-locus and four two-locus models, representing a wide range of heritability and mode of inheritance. It is of interest to see how widely we can generalize this conclusion. The CSI framework builds upon the duality of the acceptance regions of hypothesis tests and the confidence sets. More powerful tests should lead to more efficient confidence set estimators. Although the power of the traditional MLS and Mean tests, together with other linkage tests, have been extensively studied (e.g., Davis and Weeks [1997]), CSI-MLS and CSI-Mean tests use the same linkage data to test the CSI nulls that are dramatically different from the traditional linkage null. Thus, it is a new problem that warrants more investigation.

In this section we explore the power to reject the null hypothesis of complete linkage for markers with varying degrees of linkage to a disease locus. With this aim in mind, we proceed in two ways. First, we limit our attention to the three single-locus models where
we already have the comparison of efficiency of confidence sets obtained from both CSI procedures. Figure 6.2 shows the empirical power curves for CSI-MLS and CSI-Mean-V3. Both tests maintain the nominal type I error rate of 5% at the trait locus. CSI-MLS is always more powerful than CSI-Mean across the rest of the chromosome. In general, both tests attain high power under a highly heritable model, reflected in the power going quickly to 1 under model I (with heritability of 0.71). As the genetic contribution decreases (going from model I to II, and then to III), the power of both tests decrease and the power difference between the two tests becomes more pronounced. These conclusions agree with the conclusion drawn on the efficiency comparison of the two CSI procedures presented in the previous subsections.

Next, we investigate the effect of a comprehensive range of disease models on the power of the two CSI procedures to reject the CSI null at a completely informative marker that is at a range of distances away from the trait locus. The sample size was fixed at 500 ASPs. We sample 10,000 replicates for each setting, which results in the standard error of the estimate of the power being no more than 0.005. As discussed in Section 6.1.3, the effect of the disease models are entirely determined by sets of model related parameters. There are at least three sets of one-to-one parameters that play this role: (1) \((\lambda_O, \lambda_S)\), (2) \((V_A/K^2, V_D/K^2)\), and (3) \((\eta, \zeta)\). While the relative risks are appealing to epidemiologists, it is the locus specific relative risks that are relevant in our investigation. When the disease model involves more than one trait locus, this set of parameters loses its interpretational appeal. While the third set of parameters is more directly related to the power of the tests, it is far less familiar to the researchers in the field than the second set of parameters. Thus we investigate the power of the two CSI tests over a wide range of disease models as a function of \((\theta, V_A/K^2, V_D/K^2)\), \(\theta\) being the recombination fraction between the marker and the trait loci. The values of \((V_A/K^2, V_D/K^2)\) at which the powers are estimated were selected so that \(\eta < 0.65\), which covers most diseases of current interest with the trait locus contributing only a portion of the trait variability. The results are summarized in Table 6.7
and Figures 6.3 and 6.4.

From Table 6.7 and Figure 6.3, we see that the power increases with increasing distance from the trait locus. At the trait locus (0 cM), 5% of the null hypotheses are expected to be rejected, and this is satisfied up to two decimal places. At a locus that is 200 cM away from the trait locus, the power is almost 1 over the entire range of disease models. The power increases with both $V_A/K^2$ and $V_D/K^2$. Figure 6.4 shows the contour plot of the relative increase in power of CSI-MLS versus CSI-Mean at a locus that is 5 cM or 10 cM away from the trait locus. At 5 cM, CSI-MLS is almost always more powerful than CSI-Mean, with the increase in power ranging from 5% to 40%. Similar conclusions result for a marker that is 10 cM away. At large distances, both methods have power close to 1 and hence there is no visible difference in power between the two methods. Greater gains in power by using CSI-MLS instead of CSI-Mean are observed in situations where the dominance variance is the larger part of the trait variance. Such situations represent recessive diseases.

A subtlety that is not immediately apparent from Figures 6.3 and 6.4 is that, for disease models with very little genetic effect, CSI-MLS is conservative. For example, when $V_A/K^2 = 0$ and $V_D/K^2 = 0$, CSI-MLS has a type I error rate of 0% (compared to the nominal level of 5%). This is due to the non-identifiability of the likelihood when the parameters $V_A/K^2$ and $V_D/K^2$ are small. This results in the dark region near the origin in the plots of Figure 6.4. CSI-Mean maintains the type I error rates in such situations. However, no method will have any power to detect the trait locus under these disease models and hence they are of little interest. For a model with a genetic effect as modest as $\eta = 0.06$, CSI-MLS has correct type I error rates and better power than CSI-Mean (Table 6.7).
6.3 Discussion

In this Chapter, we have proposed a novel approach, CSI-MLS, that adds to the increasing literature in response to the need for construction of confidence sets of disease gene locations to better focus fine mapping efforts. The single-point CSI-MLS performs comparably with its motivator, CSI-Mean. The multi-point CSI-MLS almost always provides more precise confidence sets than CSI-Mean, over a wide range of disease models (monogenic or otherwise), marker types, and sample sizes. The only exception is the situation where the disease model is monogenic and highly heritable (with a heritability of 71%) and the map is a 7.5 cM Microsatellite one. This type of model is unrealistic for any complex trait of interest. The reduction in the lengths of the confidence sets obtained by replacing CSI-Mean with CSI-MLS, is of the order of 10-20% for many situations. There are several reasons for CSI-MLS being more efficient: (1) the MLS statistic makes maximum use of available IBD information; and (2) the MLS detects any departure from the expected IBD sharing probabilities, whereas the Mean test focuses on detecting only departure in the mean IBD sharing.

The underlying disease model is a critical factor in determining the size of the confidence sets for the trait locus. The more genetic a trait is, the shorter the interval will be. We have shown that over a large subspace of the disease models, there is considerable advantage of using CSI-MLS over CSI-Mean, with power gain as much as 40%. The gain in power is more pronounced for recessive diseases.

A second advantage of CSI-MLS over CSI-Mean is its computational efficiency. Using an AMD Athlon 2800+ (running Linux) for 100 ASPs with a 60-cM SNP map of 241 SNPs, the multi-point CSI-MLS took 2 minutes, while CSI-Mean took 31 minutes. This is not surprising given that CSI-Mean needs to perform a large number of Monte-Carlo simulations. A further advantage of CSI-MLS is that it does not need the assumption of no
missing genotypes that is required for the Monte Carlo simulations of CSI-Mean.

The SNP and MS map densities considered in the simulations are representative of current standards. At these densities, it is obvious that the SNP map leads to shorter confidence intervals than the MS maps. Our findings agree with those of Papachristou and Lin [2006a].

On the practical side, the CSI framework requires knowledge of certain disease related parameters. In this Chapter, we assume that these quantities are known. Papachristou and Lin [2006b] proposed a two-step procedure to use CSI as an intermediate mapping approach, where the required parameters were estimated in the first step. They have shown that CSI-Mean maintained the desirable statistical properties with this two-step approaches.

In the next Chapter, we find that CSI-MLS, when implemented as a two-step procedure, maintains and even enhances its advantages over CSI-Mean.

We have considered ASPs. CSI-MLS might be improved by using information from other relative types as well. Risch [1990b] showed that for diseases with large relative risks, distant relatives offer greater power to detect linkage. The likelihood in equation (6.6) can be easily generalized to other relative types. Such a generalization enables one to extract linkage information from larger pedigrees with many different relative types.

Our approach aims at providing a confidence set based on sound statistical principles and efficiency. The issue of how big a region to follow up on with fine mapping will depend also on non-statistical considerations, including the budget and the characteristics of the implicated genomic regions (such as whether this is a gene poor or gene rich region, knowledge about candidate genes, and how big the candidate genes are). Our analysis show that, for a number of disease models with small effects, CSI-MLS is able to produce a confidence region of less than 15 cMs with 500 ASPs. When one is constrained by resources or time, the proper strategy would be to start by constructing a confidence set with a suitable coverage probability that leads to a feasible region that can be studied in detail. Depending on resources, one can systematically look at the confidence sets with larger coverage proba-
Figure 6.1: The relative frequency (%) of each marker being included in the confidence sets obtained by the single-point CSI procedures. The three rows correspond to the three single-locus models in Table 6.1. A total of 30 markers in the density of one per 10 cM are considered. We set $\theta_0$ to be 5 cM in equation (6). There are 250 ASPs per replicate. The true trait locus is marked by “X”. Each tick mark represents a marker.
Figure 6.2: Power curves of CSI-MLS (solid line) and CSI-Mean-V3 (dashed line) under the three single-locus models given in Table 6.1, estimated with 1000 replicates of 250 ASPs each. The SNP map is used. Dotted line is the level of the test (5%). The trait locus is marked with $\tau^*$. 

Table 6.1: Three single-locus disease models and their parameters relevant to the power of CSI tests. 

<table>
<thead>
<tr>
<th>Model</th>
<th>$K$</th>
<th>$\lambda_O$</th>
<th>$\lambda_S$</th>
<th>$H = \frac{V_G^a}{V_T}$</th>
<th>$\eta$</th>
<th>$\zeta$</th>
<th>$P_D^b$</th>
<th>$f_{DD}^c$</th>
<th>$f_{Dd}^c$</th>
<th>$f_{dd}^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$3.4950 \times 10^{-3}$</td>
<td>10.68</td>
<td>56.68</td>
<td>0.71</td>
<td>0.982</td>
<td>0.812</td>
<td>0.050</td>
<td>0.999</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>II</td>
<td>$1.0000 \times 10^{-1}$</td>
<td>2.25</td>
<td>2.75</td>
<td>0.50</td>
<td>0.636</td>
<td>0.182</td>
<td>0.226</td>
<td>0.998</td>
<td>0.105</td>
<td>0.020</td>
</tr>
<tr>
<td>III</td>
<td>$8.4625 \times 10^{-2}$</td>
<td>1.78</td>
<td>1.78</td>
<td>0.14</td>
<td>0.438</td>
<td>0.001</td>
<td>0.010</td>
<td>0.865</td>
<td>0.827</td>
<td>0.070</td>
</tr>
</tbody>
</table>

$^a$ $V_G$: genetic variance of the trait; $V_T$: total variance of the trait; $H$: broad sense heritability.

$^b$ $D$: High risk allele; $d$: low risk allele; $P_D$: allele frequency of $D$.

$^c$ $(f_{DD}, f_{Dd}, f_{dd})$: penetrances of the genotypes $(DD, Dd, dd)$.
Figure 6.3: Plots of the power of CSI-MLS and CSI-Mean at different distances from the true disease locus for a range of disease values represented by the parameters $(V_A/K^2, V_D/K^2)$.

Table 6.2: Two-locus disease models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$f_{22}$</th>
<th>$f_{21}$</th>
<th>$f_{20}$</th>
<th>$f_{12}$</th>
<th>$f_{11}$</th>
<th>$f_{10}$</th>
<th>$f_{02}$</th>
<th>$f_{01}$</th>
<th>$f_{00}$</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP-2</td>
<td>$\phi$</td>
<td>$\phi$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.600</td>
<td>.199</td>
<td>.778</td>
</tr>
<tr>
<td>EP-4</td>
<td>$\phi$</td>
<td>$\phi$</td>
<td>0</td>
<td>$\phi$</td>
<td>0</td>
<td>$\phi$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.372</td>
<td>.243</td>
<td>.911</td>
</tr>
<tr>
<td>Het-2</td>
<td>$\varphi$</td>
<td>$\varphi$</td>
<td>$\phi$</td>
<td>$\phi$</td>
<td>0</td>
<td>$\phi$</td>
<td>0</td>
<td>$\phi$</td>
<td>0</td>
<td>.279</td>
<td>.040</td>
<td>.660</td>
</tr>
<tr>
<td>S-2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$\phi$</td>
<td>$\phi$</td>
<td>0</td>
<td>$\phi$</td>
<td>0</td>
<td>$\phi$</td>
<td>.228</td>
<td>.045</td>
<td>.574</td>
</tr>
</tbody>
</table>

$^a$ $f_{ij}$ ($i, j = 0, 1, 2$): the penetrance of a genotype with $i$ copies of high risk allele at the first locus and $j$ copies of high risk allele at the second locus.

$^b$ $p_i$ ($i = 1, 2$): frequency of the high risk allele at the $i^{th}$ trait locus.

$^c$ $\varphi = 2\phi - \phi^2$: the penetrance of a genotype with the high risk genotypes at each of the two component trait loci, under a heterogeneity model.
Figure 6.4: Contour plots of relative differences (%) in power in terms of \( \frac{V_A}{K^2}, \frac{V_D}{K^2} \), at a marker that is 5 cM or 10 cM away from the trait locus. The relative difference is defined as \( \frac{\text{Power of CSI-MLS} - \text{Power of CSI-Mean}}{\text{Power of CSI-Mean}} \times 100\% \).

Table 6.3: Parameters relevant to the power of the CSI tests for the two-locus trait models in Table 6.2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Trait locus I ( ^a )</th>
<th>Trait locus II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( z_0 )</td>
<td>( z_1 )</td>
</tr>
<tr>
<td>EP-2</td>
<td>0.141</td>
<td>0.469</td>
</tr>
<tr>
<td>EP-4</td>
<td>0.194</td>
<td>0.467</td>
</tr>
<tr>
<td>Het-2</td>
<td>0.175</td>
<td>0.441</td>
</tr>
<tr>
<td>S-2</td>
<td>0.146</td>
<td>0.405</td>
</tr>
</tbody>
</table>

\( ^a z_k (k = 0, 1, 2): \) marginal IBD probability for an ASP at the trait locus under consideration; \( \lambda_O \) and \( \lambda_S \): locus specific relative risks to an offspring and to a sibling of an affected individual, respectively.

87
Table 6.4: Confidence sets from the three multi-point CSI procedures on the MS map. Information includes mean length of confidence sets in cM (M), standard deviation of the lengths (SD), and coverage probability (CP).

<table>
<thead>
<tr>
<th>Model</th>
<th>No. ASPs</th>
<th>CSI-MLS M</th>
<th>SD</th>
<th>CP</th>
<th>CSI-Mean-V3 M</th>
<th>SD</th>
<th>CP</th>
<th>CSI-Mean-V2 M</th>
<th>SD</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>10.6</td>
<td>3.5</td>
<td>0.945</td>
<td></td>
<td></td>
<td>10.6</td>
<td>2.7</td>
<td>0.949</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>250</td>
<td>8.8</td>
<td>3.3</td>
<td>0.940</td>
<td></td>
<td></td>
<td>8.8</td>
<td>2.1</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>7.7</td>
<td>3.0</td>
<td>0.943</td>
<td></td>
<td></td>
<td>7.8</td>
<td>1.8</td>
<td>0.932</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>14.9</td>
<td>6.4</td>
<td>0.944</td>
<td></td>
<td></td>
<td>17.4</td>
<td>6.2</td>
<td>0.950</td>
</tr>
<tr>
<td>II</td>
<td></td>
<td>100</td>
<td>25.4</td>
<td>11.8</td>
<td>0.942</td>
<td></td>
<td></td>
<td>28.7</td>
<td>12.0</td>
<td>0.941</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>16.5</td>
<td>6.5</td>
<td>0.957</td>
<td></td>
<td></td>
<td>18.7</td>
<td>6.4</td>
<td>0.953</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>13.0</td>
<td>4.7</td>
<td>0.953</td>
<td></td>
<td></td>
<td>14.3</td>
<td>4.5</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>14.9</td>
<td>6.4</td>
<td>0.944</td>
<td></td>
<td></td>
<td>17.4</td>
<td>6.2</td>
<td>0.950</td>
</tr>
<tr>
<td>III</td>
<td></td>
<td>100</td>
<td>29.5</td>
<td>14.4</td>
<td>0.931</td>
<td></td>
<td></td>
<td>36.0</td>
<td>15.2</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>20.8</td>
<td>9.5</td>
<td>0.941</td>
<td></td>
<td></td>
<td>24.6</td>
<td>9.6</td>
<td>0.931</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>14.9</td>
<td>6.4</td>
<td>0.944</td>
<td></td>
<td></td>
<td>17.4</td>
<td>6.2</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>14.9</td>
<td>6.4</td>
<td>0.944</td>
<td></td>
<td></td>
<td>17.4</td>
<td>6.2</td>
<td>0.950</td>
</tr>
</tbody>
</table>
Table 6.5: Confidence sets from multi-point CSI procedures on the SNP map. Information includes mean length of confidence sets in cM (M), effect of sample size (SS: \(\frac{\text{length of CI}}{\text{length of CI with 100 ASPs}} \times 100\%\)), and relative reduction in length obtained by using CSI-MLS instead of CSI-Mean (\(\text{RR} = \frac{\text{Length of CSI-Mean} - \text{Length of CSI-MLS}}{\text{Length of CSI-Mean}} \times 100\%\)).

<table>
<thead>
<tr>
<th>Model</th>
<th>No. ASPs</th>
<th>CSI-MLS</th>
<th>CSI-Mean-V3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SS (%)</td>
<td>M</td>
</tr>
<tr>
<td>I</td>
<td>100</td>
<td>4.2</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>2.6</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.9</td>
<td>45</td>
</tr>
<tr>
<td>II</td>
<td>100</td>
<td>18.5</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>10.5</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>7.3</td>
<td>39</td>
</tr>
<tr>
<td>III</td>
<td>100</td>
<td>35.0</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>21.2</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>13.7</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>9.1</td>
<td>26</td>
</tr>
</tbody>
</table>
Table 6.6: Confidence sets from CSI procedures on two-locus disease models. Information includes mean length of confidence sets in cM (M), effect of sample size (SS: \(\frac{\text{length of CI}}{\text{length of CI with 100 ASPs}} \times 100\%\)), and relative reduction in length obtained by using CSI-MLS instead of CSI-Mean (RR: \(\frac{\text{Length of CSI-Mean} - \text{Length of CSI-MLS}}{\text{Length of CSI-Mean}} \times 100\%\)).

<table>
<thead>
<tr>
<th>Model</th>
<th>No. ASPs</th>
<th>Trait Locus 1</th>
<th></th>
<th></th>
<th>Trait Locus 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CSI-MLS M</td>
<td>SS</td>
<td>CSI-Mean-V3 M</td>
<td>SS</td>
<td>RR</td>
<td>CSI-MLS M</td>
</tr>
<tr>
<td>EP-2</td>
<td>100</td>
<td>48.1 –</td>
<td>–</td>
<td>57.8 –</td>
<td>–</td>
<td>16.8</td>
<td>50.1 –</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>18.7 39</td>
<td></td>
<td>21.5 37</td>
<td></td>
<td>13.0</td>
<td>19.4 39</td>
</tr>
<tr>
<td>EP-4</td>
<td>100</td>
<td>78.1 –</td>
<td>–</td>
<td>91.5 –</td>
<td>–</td>
<td>23.0</td>
<td>29.3 –</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>34.2 44</td>
<td></td>
<td>44.1 48</td>
<td></td>
<td>22.4</td>
<td>13.3 45</td>
</tr>
<tr>
<td>Het-2</td>
<td>100</td>
<td>59.3 –</td>
<td>–</td>
<td>74.8 –</td>
<td>–</td>
<td>20.1</td>
<td>58.3 –</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>22.1 37</td>
<td></td>
<td>27.4 37</td>
<td></td>
<td>19.3</td>
<td>21.4 36</td>
</tr>
<tr>
<td>S-2</td>
<td>100</td>
<td>36.9 –</td>
<td>–</td>
<td>48.1 –</td>
<td>–</td>
<td>23.3</td>
<td>74.9 –</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>16.2 44</td>
<td></td>
<td>19.6 41</td>
<td></td>
<td>17.3</td>
<td>29.5 39</td>
</tr>
</tbody>
</table>
Table 6.7: Power of CSI-MLS and CSI-Mean.

<table>
<thead>
<tr>
<th>$\theta$ (cM)</th>
<th>$V_A/K^2$</th>
<th></th>
<th></th>
<th>$V_D/K^2$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.12</td>
<td>0.36</td>
<td>0.72</td>
<td>1.08</td>
<td>1.76</td>
<td>2.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.06</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.54</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.06, 0.05</td>
<td>0.05, 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.44</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td>0.05, 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.06</td>
<td>0.09, 0.08</td>
<td>0.19, 0.14</td>
<td>0.37, 0.22</td>
<td>0.55, 0.33</td>
<td>0.81, 0.52</td>
<td>0.96, 0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>0.11, 0.11</td>
<td>0.20, 0.15</td>
<td>0.38, 0.25</td>
<td>0.55, 0.35</td>
<td>0.80, 0.54</td>
<td>0.96, 0.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>0.15, 0.14</td>
<td>0.23, 0.21</td>
<td>0.38, 0.29</td>
<td>0.54, 0.40</td>
<td>0.77, 0.56</td>
<td>0.95, 0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.54</td>
<td>0.19, 0.18</td>
<td>0.26, 0.24</td>
<td>0.40, 0.35</td>
<td>0.54, 0.44</td>
<td>0.75, 0.59</td>
<td>0.94, 0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.26, 0.26</td>
<td>0.32, 0.32</td>
<td>0.44, 0.41</td>
<td>0.56, 0.50</td>
<td>0.74, 0.64</td>
<td>0.92, 0.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.44</td>
<td>0.39, 0.38</td>
<td>0.44, 0.44</td>
<td>0.51, 0.50</td>
<td>0.60, 0.60</td>
<td>0.75, 0.72</td>
<td>0.90, 0.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.06</td>
<td>0.13, 0.11</td>
<td>0.37, 0.25</td>
<td>0.74, 0.49</td>
<td>0.93, 0.71</td>
<td>1, 0.93</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>0.18, 0.17</td>
<td>0.42, 0.31</td>
<td>0.75, 0.56</td>
<td>0.92, 0.76</td>
<td>1, 0.95</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>0.28, 0.27</td>
<td>0.48, 0.43</td>
<td>0.77, 0.65</td>
<td>0.92, 0.82</td>
<td>0.99, 0.96</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.54</td>
<td>0.38, 0.38</td>
<td>0.57, 0.53</td>
<td>0.80, 0.72</td>
<td>0.93, 0.86</td>
<td>0.99, 0.97</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.56, 0.56</td>
<td>0.69, 0.69</td>
<td>0.85, 0.82</td>
<td>0.94, 0.91</td>
<td>0.99, 0.98</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.44</td>
<td>0.79, 0.78</td>
<td>0.85, 0.85</td>
<td>0.91, 0.91</td>
<td>0.96, 0.96</td>
<td>0.99, 0.99</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.06</td>
<td>0.28, 0.23</td>
<td>0.81, 0.65</td>
<td>1, 0.97</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>0.46, 0.44</td>
<td>0.88, 0.80</td>
<td>1, 0.98</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>0.71, 0.70</td>
<td>0.95, 0.93</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.54</td>
<td>0.88, 0.88</td>
<td>0.98, 0.98</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.99, 0.98</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.44</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.06</td>
<td>0.42, 0.36</td>
<td>0.95, 0.90</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>0.70, 0.68</td>
<td>0.98, 0.97</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>0.93, 0.93</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.54</td>
<td>0.99, 0.99</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.44</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1, 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Each cell gives (probability of rejection for CSI-MLS, probability of rejection for CSI-Mean).

$^b$ Each value is correct to 2 decimal places with a standard error of at most 0.005.
Chapter 7

Two-Step Intermediate Fine Mapping with Likelihood Ratio Test Statistics

With the advent of high-throughput genotyping technologies, traditional gene mapping methods, including linkage approaches, may be improved to realize the full potential of the wealth of genotype data available today. To better localize disease genes using today's dense maps, Confidence Set Inference (CSI) constructs precise confidence sets for disease gene locations [Papachristou and Lin, 2006a] using affected sibling pair (ASP) data. In addition to guiding the fine mapping efforts following linkage analysis, CSI enhances objectivity by providing confidence sets that contain the disease gene location with a predetermined probability. The original CSI framework was implemented using the Mean test statistic (CSI-Mean) [Papachristou and Lin, 2006a]. In the last Chapter, we proposed a more efficient alternative, CSI-MLS, by reformulating the Maximum LOD Score (MLS) statistic [Risch, 1990c], a Likelihood Ratio Test (LRT) statistic. The CSI procedures require knowledge of some disease model related parameters. When these parameters are known, CSI-MLS has substantial advantages over CSI-Mean: (1) it provides more precise confidence sets for disease gene locations with correct coverage probabilities; (2) it is computationally more efficient. In the absence of knowledge of these parameters, a two-step
procedure that constructs the confidence sets in the second step following the estimation of the required parameters in the first step may be adopted [Papachristou and Lin, 2006b]. In this Chapter we show that the advantages of CSI-MLS are preserved in the two-step procedures. Also investigated are the effects of three different strategies in step one and the impact of the absence of parental genotype data.

### 7.1 Methods

A multi-point CSI framework for construction of confidence sets for disease gene locations was proposed recently [Papachristou and Lin, 2006a]. In this framework, each genomic location \((\tau)\) in a broad region with linkage evidence is tested to see whether it is the putative disease causing locus \((\tau^*)\). In contrast to the traditional null hypothesis of no linkage, a new hypothesis is tested:

\[
H_0_{\tau} : \tau = \tau^* \text{ versus } H_{A\tau} : \tau \neq \tau^*.
\] (7.1)

The confidence set for the disease location contains all the genomic locations where the above null hypothesis is not rejected. Any linkage statistic can be reformulated to test hypothesis (7.1), including the Mean Test statistic (CSI-Mean, [Papachristou and Lin, 2006a]) and the LRT statistic (CSI-MLS, Chapter 6). Both procedures assume knowledge of some disease model related parameters, one possibility being \(z_i = P(\tau^* IBD = i), i = 0, 1, 2,\) where \(\tau^* IBD\) is the identical by descent (IBD) allele sharing between two affected siblings at the disease locus. CSI-MLS provides more precise confidence sets compared to CSI-Mean when \(z_i\)'s are known (Chapter 6).

In the realistic scenario when these parameters are not known \textit{a priori}, a two-step procedure that relies on the availability of two sets of marker data on the same set of ASPs can be adopted [Papachristou and Lin, 2006b]: (1) Identify linkage regions using one set of
markers, termed the coarse map (e.g. microsatellite markers). Genomic locations with a nonparametric statistic (KAC) [Kong and Cox, 1997] above 2.33 are identified as showing suggestive linkage. For the linkage peak in each region that exceeds the above cut-off, obtain maximum likelihood estimates of the parameters $z_i$ [Risch, 1990c]. The only assumption required is that there is only one disease predisposing loci on the chromosome in question. (2) Using a set of denser markers, termed the fine map (e.g. SNP markers), and restricted to a region of, say, 25 cM on either side of the linkage peak, we construct confidence sets employing CSI-MLS or CSI-Mean with the parameter estimates from step (1). Since confidence sets so obtained may not be contiguous, we employed the smoothing scheme suggested by [Papachristou and Lin, 2006b].

Traditionally, gene mapping is often pursued via a linkage genome scan with microsatellite markers, and then the preliminary linkage signals are followed up using a denser, usually SNP, map. It was this practice that motivated the two-step CSI procedures. However, with the advent of high density genomewide SNP chips, this practice may soon be replaced by a single dense SNP map on which all individuals are genotyped. The 5K and 10K chips have already been successfully employed in linkage analysis. In such situations, we propose using a subset of the SNP markers as the coarse map and another mutually exclusive subset of the SNP markers as the fine map. Within each of these two maps, markers are chosen to be in linkage equilibrium with each other ($r^2 < 0.02$), as linkage disequilibrium between markers lead to erroneous estimates of multi-point IBDs [Xing et al., 2006].

We apply our methods to Genetic Analysis Workshop 15 (http://www.gaworkshop.org/) simulated data set, with the phenotype being the binary trait of the affection status of Rheumatoid Arthritis (RA). The microsatellite and SNP data (excluding the dense SNPs for the simulated data) on Chromosome 6, containing the DRB1 locus of HLA, are used. The microsatellite and SNP maps were compared in terms of how well the disease model related parameters are estimated, and the two-step CSI-MLS and CSI-Mean were com-
pared in terms of the precision of the confidence sets that include the DRB1 locus with a pre-determined probability. In what follows, we consider all of the 100 replicates, with a sample size of \( n \) (250, 500, or 750) per replicate taken to be the first \( n \) families of each replicate. Estimates of means, standard deviations, etc., of the parameter of interest are based on the 100 replicates of the respective sample size. The simulating model was known in advance.

### 7.2 Results

Central to the CSI framework is the knowledge of the IBD probabilities at the trait locus, \( z_0, z_1 \), and \( z_2 \). While this information is not available from the “Answers” to the simulated data, a good guess is the MLE at the HLA-DRB1 locus obtained from the chromosome 6 SNP data on all families (with parental genotypes) from all available replicates. The MLEs, \(( \hat{z}_0 = 0.101, \hat{z}_1 = 0.441, \hat{z}_2 = 0.458 )\), will be considered to be the true values of the corresponding parameters. The information content [Kruglyak and Lander, 1995] at the trait locus was 0.97, which suggests that these estimates are very close to the real value. In what follows, we investigate the effect of the coarse map on the step 1 parameter estimates, followed by a comparison of the performance of CSI-MLS and CSI-Mean in the two-step setting. We also study the influence of the availability of parental genotype information.

#### 7.2.1 Estimates of \(( z_0, z_1, z_2 )\) at the disease locus

Three estimation strategies/settings have been explored in step 1 of CSI: (1) the MLE is evaluated at the microsatellite marker with the largest KAC score (MS); (2) the MLE is evaluated at the location, possibly between two microsatellite markers, with the highest KAC score (MSINT); (3) the MLE is evaluated with a SNP map as the coarse map. The
MS and SNP coarse maps had 41 microsatellites and 292 SNPs, respectively. With parental data, the information content at the disease locus was 0.85 for the MS map and 0.88 for the SNP map; without parental data, it was 0.56 for both. The precision of the parameter estimates were evaluated in terms of Mean Squared Errors (MSEs). Table 7.1 gives the MSEs of the estimates of the disease gene locations (peak of KAC scores) and of the $z_i$s, with and without parental genotypes. All estimates improve with increasing sample size. With parental genotypes, SNP yields best estimate of $z_i$s, with MSINT coming next and MS being the least favorable, though there is no pronounced difference. This finding agrees well with the information content of the maps. However, when parental genotypes are not available, MS yields much more precise estimates than SNP and MSINT. A similar trend is reflected in Figure 7.1. When parental genotypes are missing, there seems to be a systematic bias in the estimates under all three setting, $z_1$ being underestimated (for 750 ASPs, the bias was -.01, -.03 and -.03 for MS, MSINT and SNP, respectively) and $z_2$ being overestimated (for 750 ASPs, the bias was .01, .03 and .02 for MS, MSINT and SNP, respectively). As is seen in the next Section this observed bias does not adversely affect the CSI-MLS and CSI-Mean procedures.

Cordell [2004] performed a similar experiment. The author concluded that the MLE’s of IBD sharing probabilities at the disease location showed no bias even when some of the ASPs are not informative for number of alleles shared IBD (both parents homozygous for same allele). However, we see that when parental genotypes are missing the MLE’s of $z_i$’s are indeed biased. This contradiction may be explained by the following differences in simulation settings. First of all, the bias in the estimates in our situation is only observed when parental genotypes are missing. Cordell [2004] simulated a marker at the disease locus and two more markers 100 cM from the disease locus on either side. Apart from the proportion of ASPs not informative for linkage at the disease locus, the markers were completely informative. This is also drastically different from our situation where markers are more densely spaced. The information about IBD sharing is obtained in a multi-point fashion in
our situation (adjacent markers providing information), where as in the simulation settings of Cordell [2004] most of the IBD sharing information is given by the completely informative marker. In light of our findings, if unbiased estimation of the $z_i$ is the objective, one should not use dense maps in pedigrees with missing parental genotypes.

### 7.2.2 Comparison of CSI-MLS and CSI-Mean

Using the estimates of the $z_i$ from step 1, CSI-Mean and CSI-MLS confidence sets are constructed with a fine map of 350 SNP markers on the chromosome. Besides the three settings MS, MSINT, and SNP, we added the group TRUE, where the CSI confidence sets were constructed assuming the true parameter values. Mean lengths of the confidence sets, together with the empirical coverage levels, are given in Table 7.2. The coverage levels under TRUE are close to the nominal 95%, while they are higher than the nominal levels in all other situations. We compare the precision of the confidence sets in terms of their lengths, since the coverage is at least up to the nominal level. Under each setting, the confidence sets get tighter with increased sample size. CSI-MLS provides substantially shorter confidence sets than CSI-Mean, the effect being more pronounced when parental genotypes are not available and when true parameter estimates are not used. Among the two-step procedures, in all situations MSINT provides the most precise confidence sets, with SNP performing only slightly worse. TRUE performs the best when parental genotypes are known, but results in longer confidence sets than the other settings without parental genotypes.

### 7.3 Discussion

We demonstrate here that CSI-MLS performs substantially better than CSI-Mean, with the advantage more apparent when parental genotype data are missing. This is not surprising
as the trait model represents a favorable situation for CSI-MLS (Chapter 6). In a disease with no dominance genetic variance the value of $z_1$ is 0.5. The value of $z_1$ for RA at the HLA-DRB1 locus is 0.441, significantly different from 0.5. This difference suggests a reasonable contribution of dominance genetic variance to the total variance. CSI-MLS performs better than CSI-Mean in most situations, with the gain most substantial when the disease has dominance variance (Chapter 6). The better use of allelic IBD information by the MLS statistic might explain the more pronounced advantage of CSI-MLS in the absence of parental genotype data.

When parental genotype information is not available, parameter estimates at microsatellite markers appeared to be superior to MSINT and SNP. However, when the precision of the confidence sets are considered, MSINT provided the shortest sets while maintaining the nominal coverage probabilities. This apparent contradiction could be explained as follows. When parental information is not available, there is a bias in the estimates of $z_i$'s under both MSINT and SNP settings. However, parental data is also absent on the fine map, which leads to a similar bias in step two. As long as these two biases are consistent (in the same direction), which is reasonable, we believe the coverage probabilities will not be affected. Another striking observation is that CSI-MLS confidence sets are shorter when parental genotypes are missing. This is because the empirical coverage level is smaller, albeit still above 95%, when the parental genotypes are missing. Our suggestion on what method to use for CSI is: when microsatellite and SNP genotypes are available on the same set of families, use MSINT to estimate the parameters in step one and then construct the confidence sets using CSI-MLS on a dense SNP map in step two. When microsatellite data are not available, use two mutually exclusive dense SNP maps in the two steps. Even though the use of SNPs in step one leads to some loss in efficiency, this loss is limited. This loss can presumably be reduced by using a denser SNP map in step one, though one has to be careful about the LD between markers affecting parameter estimates.
Figure 7.1: Density of the estimates of $z_1$ and $z_2$, with and without parental genotypes, using 750 ASPs. The vertical line represents the true value of the parameter.

Table 7.1: Mean Squared Errors of the estimates of the disease gene locations ($\tau^*$ in cM) and the IBD probabilities ($z_0, z_1, z_2$), using the strategies MS, MSINT, and SNP. The entries under the $z$’s are (MSE $\times 10^4$).

<table>
<thead>
<tr>
<th># ASP</th>
<th>Strategy</th>
<th>With Parental Genotypes</th>
<th>Without Parental Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau^*$</td>
<td>$z_0$</td>
<td>$z_1$</td>
</tr>
<tr>
<td>250</td>
<td>MS</td>
<td>23.65</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>MSINT</td>
<td>14.10</td>
<td>4.01</td>
</tr>
<tr>
<td></td>
<td>SNP</td>
<td>6.00</td>
<td>3.40</td>
</tr>
<tr>
<td>500</td>
<td>MS</td>
<td>13.79</td>
<td>2.62</td>
</tr>
<tr>
<td></td>
<td>MSINT</td>
<td>3.79</td>
<td>2.46</td>
</tr>
<tr>
<td></td>
<td>SNP</td>
<td>1.76</td>
<td>2.00</td>
</tr>
<tr>
<td>750</td>
<td>MS</td>
<td>13.27</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>MSINT</td>
<td>2.13</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>SNP</td>
<td>1.64</td>
<td>1.52</td>
</tr>
</tbody>
</table>
Table 7.2: Mean lengths of 95% confidence sets obtained using CSI-MLS (MLS) and CSI-Mean (Mean). Empirical proportions of confidence sets that contain the disease locus are given in parentheses. RR, defined as \( \frac{\text{Mean length of CSI-MLS} - \text{Mean length of CSI-Mean}}{\text{Mean length of CSI-Mean}} \times 100\% \), is the relative reduction in length obtained by using CSI-MLS in place of CSI-Mean.

<table>
<thead>
<tr>
<th># ASP</th>
<th>Strategy</th>
<th>With Parental Genotypes</th>
<th>Without Parental Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MLS</td>
<td>Mean</td>
</tr>
<tr>
<td>250</td>
<td>MS</td>
<td>28.1</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>MSINT</td>
<td>22.8</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>SNP</td>
<td>24.6</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>TRUE</td>
<td>21.5</td>
<td>(.91)</td>
</tr>
<tr>
<td>500</td>
<td>MS</td>
<td>21.8</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>MSINT</td>
<td>15.4</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>SNP</td>
<td>17.0</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>TRUE</td>
<td>15.2</td>
<td>(.92)</td>
</tr>
<tr>
<td>750</td>
<td>MS</td>
<td>18.4</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>MSINT</td>
<td>12.2</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>SNP</td>
<td>14.2</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>TRUE</td>
<td>11.9</td>
<td>(.94)</td>
</tr>
</tbody>
</table>
Chapter 8

Bayesian Intervals for Linkage Location

Confidence sets for the disease gene location constructed after initial linkage region identification can help narrow the region in which fine mapping studies are conducted. Along with reduction in the cost of follow-up studies, they can reduce the multiple testing associated with fine mapping studies. The Confidence Set Inference (CSI) provides a promising framework for constructing confidence sets with controlled coverage probabilities. However, one drawback with the CSI methods is that, in most real situations, the requirement of knowledge of IBD sharing probabilities at the disease locus, is not satisfied. Papachristou and Lin [2006b] proposed a Two-Step procedure to overcome this shortcoming. This Two-Step procedure makes CSI easily applicable to most real affected sibling pair genotype data. The Two-Step procedure estimates the required disease model related parameters from a map in step one (e.g. a Microsatellite map) and uses these estimates to construct CSI confidence sets in a distinct map (e.g. a SNP map) in the second step. The Two-Step CSI procedure has been observed to have greater than nominal coverage of the disease locus and hence the intervals are sometimes longer than desirable (e.g., Chapter 7, Papachristou and Lin [2006c]).
Papachristou and Lin [2006c] compared five intermediate fine mapping methods, including, CSI with Mean test statistic (CSI-Mean), LOD support intervals, GEE based confidence sets, and two proposed bootstrap methods. While the LOD support intervals and the GEE based methods were observed to have reduced coverage (drastically reduced for small sample sizes), CSI-Mean had greater than nominal coverage. The two proposed bootstrap methods had close to nominal coverage, but were not the shortest when the empirical coverage were held to the same level. In Chapter 6 we proposed testing the CSI hypothesis using a likelihood ratio test (CSI-MLS).

Risch [1990c] proposed a likelihood for ASP linkage data with three free parameters $l$, the location of the disease gene, $(z_0, z_1, z_2)$, the probabilities of sharing 0, 1, or 2 alleles IBD at the disease locus by an ASP (with $z_0 + z_1 + z_2 = 1$). The maximization of this likelihood, under the triangle constraints ($z_1 \leq 1/2$ and $2z_0 \leq z_1$) of Holmans [1993], provides a consistent estimate of the location of the putative gene [Risch, 1990c, Holmans, 1993, Cordell, 2004, Papachristou and Lin, 2006c]. This maximum likelihood estimator (MLE) may be used to provide asymptotic confidence intervals for the parameter. However, there are a few issues with this approach. First, construction of asymptotic confidence intervals will require knowledge of large sample variance of these parameter estimates [Louis, 1982]. But recall that Holmans [1993] used a modified Expectation Maximization algorithm. A second problem with this approach will be the violation of the assumption of normality when the true disease gene is in the telomeric region. Finally, as pointed out by Papachristou and Lin [2006c], the maximization of the likelihood is conducted under the triangle constraints, and such a constrained maximization will mean that the asymptotic normality of the MLE may be violated. In order to estimate the variance of the parameter estimate, Papachristou and Lin [2006c] proposed two Bootstrap procedures. A Bootstrap procedure often leads to poorer estimates of a parameter than when knowledge about the problem at hand is incorporated into the statistical model. Choosing such a procedure will involve compromising efficiency for robustness.
In this Chapter, we propose a Bayesian procedure to estimate the location of the disease causing gene. Given a prior probability for the location of the disease gene, a set having a posterior probability of \((1 - \epsilon)\) is called a Bayesian credible interval. As discussed in Dupuis and Siegmund [1999], \((1 - \epsilon)\) Bayesian credible sets, with a uniform prior on the location, are in fact \((1 - \epsilon)\) confidence regions having many desirable properties. We propose using these credible intervals as interval estimators of the parameter. The primary parameter of interest is the underlying disease gene location, \(l\). The likelihood involves three parameters, in addition to \(l\), the additional parameters are \(\alpha = V_A/K(1 - K)\) and \(\delta = V_D/K(1 - K)\), where \(K\) is the population prevalence of the disease, \(V_A\) is the additive genetic variance, and \(V_D\) is the dominance genetic variance. These parameters have natural genetic interpretations, discussed later. These two parameters along with \(K\) may be re-parameterized as \((z_0, z_1, z_2)\).

### 8.1 Methods

#### 8.1.1 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. The state of the chain after a large number of steps is then used as a sample from the desired distribution. The Metropolis-Hastings algorithm [Gilks et al., 1996] provides a way of sampling from the target distribution, which in the case of a Bayesian problem is the posterior distribution of the parameters given the data. Let \(\theta\) and \(x\) denote vectors of parameters and data, respectively. And let the posterior distribution of interest be given by \(\pi(\theta|x)\). Then the Metropolis-Hastings algorithm to sample from \(\pi(\theta|x)\) is given by the following steps. Begin the chain with an arbitrary starting value \(\theta^{(0)}\), and then produce the chain \(\theta^{(1)}, \theta^{(2)}, \ldots\) by iterating with the following two steps.
At iteration $i + 1$: (1) generate candidate value $\theta'$ from the proposal distribution $q(\cdot|\theta^{(i)})$; 
(2) With probability 

$$A(\theta^{(i)}, \theta') = \min \left[ 1, \frac{\pi(\theta'|x)q(\theta^{(i)}|\theta')}{\pi(\theta^{(i)}|x)q(\theta'|\theta^{(i)})} \right]$$

accept the candidate, otherwise reject the candidate. The posterior distribution $\pi(\theta|x)$ is proportional to the likelihood $P(x|\theta)$ times the prior $\pi(\theta)$, where the likelihood $P(x|\theta)$ only needs to be known up to a multiplicative constant. Estimates of parameters of interest, say $E_{\pi}(g(\theta)|x)$, are obtained by summing over samples from the chain after the chain is believed to have converged to the stationary distribution. MCMC methods are useful in numerous situations because parameters of interest can often be framed as means.

### 8.1.2 Likelihood

Let $G_i$ denote the genotypes of all markers on the chromosome under consideration for the $i^{th}$ ASP (and possibly family members), $i = 1, \ldots, n$. Also, define $z_j = P(I\!BD = j|\text{Both sibs are affected})$, $j = 0, 1, 2$, $l$ denotes the true underlying disease location and $tI\!BD_i$ be the number of alleles shared identical by descent by ASP $i$ at the location $t$. The
likelihood function for all $n$ ASPs is given by

$$L(l, \alpha, \delta; \text{Data}) = \prod_{i=1}^{n} P(G_i | l, \alpha, \delta)$$

$$= \prod_{i=1}^{n} \sum_{j=0}^{2} P(G_i | lIBD_i = j | l, \alpha, \delta)$$

$$= \prod_{i=1}^{n} \sum_{j=0}^{2} P(G_i | lIBD_i = j, l, \alpha, \delta)P(lIBD_i = j | l, \alpha, \delta)$$

$$= \prod_{i=1}^{n} \sum_{j=0}^{2} P(G_i | lIBD_i = j)P(lIBD_i = j | l, \alpha, \delta)$$

$$= \prod_{i=1}^{n} P(G_i) \sum_{j=0}^{2} \frac{P(lIBD_i = j | G_i)}{P(lIBD_i = j)} z_j(\alpha, \delta), \quad (8.1)$$

where, for a single locus disease model $z_j(\alpha, \delta)$ is given by [James, 1971]

$$z_0(\alpha, \delta) = \frac{K/(1-K)}{4K/(1-K) + 2\alpha + \delta}$$

$$z_1(\alpha, \delta) = \frac{2K/(1-K) + \alpha}{4K/(1-K) + 2\alpha + \delta}$$

$$z_2(\alpha, \delta) = \frac{K/(1-K) + \alpha + \delta}{4K/(1-K) + 2\alpha + \delta}.$$

We assume that $K$, the prevalence, is known without uncertainty. This assumption is necessary because the sample is ascertained to include affected sibling pairs, and the likelihood provides no information to estimate $K$. Prevalence of a disease is easily estimated from random epidemiologic samples from the population. In a homogeneous population this assumption should well satisfied. The above likelihood is known only up to a constant because of the computational difficulty in calculating the probabilities $P(G_i)$. The probabilities $P(lIBD_j = i)$ are simply the traditional null IBD sharing probabilities, for siblings they are (1/4, 1/2, 1/4) for sharing (0, 1, 2) alleles IBD, respectively. Finally, the probability $P(lIBD_i = j | G_i)$ depends on the observed genotypes of the $j^{th}$ family. These probabilities are well known quantities [Haseman and Elston, 1972] and are provided by a number of
genetic software (e.g., S.A.G.E. [2006], Merlin [Abecasis et al., 2002] and Allegro [Gudbjartsson et al., 2000]). The above likelihood requires computation of the IBD sharing probabilities given the data at the location \( l \); because of this computational requirement, \( l \) was assumed to take values on a fine grid (of say every 0.25 cM on the chromosome).

The lack of knowledge of \( P(G_i) \) means that the full conditional distribution of the parameters is not known, and as a result a Gibbs sampler cannot be conducted [Gilks et al., 1996]. Hence, a Markov Chain Monte Carlo (MCMC) simulation using the single component Metropolis-Hastings algorithm is run to sample from the posterior distribution.

### 8.1.3 Prior Distributions

The parameter \( l \) was chosen to have a uniform prior on \( \{0, d, 2d, \ldots, L\} \), where \( L \) is the location in cM of the last marker on the chromosome, and \( d \) is distance between adjacent locations making up the grid. The parameter \( l \) must have a uniform prior for the resulting credible interval to be interpretable as a confidence interval. The prior distribution of \( l \) is assumed to be independent of the other two parameters, the priors of which are specified in the next paragraph.

The prior distribution of \((\alpha, \delta)\) is motivated as follows. Let \( Y \) be the random variable indicating affection by a genetic disease. That is,

\[
Y = \begin{cases} 
1 & \text{if the individual is affected}, \\
0 & \text{if the individual is unaffected}.
\end{cases}
\]

Then, \( E(Y) = K \), the prevalence of the disease. The total variance of the trait is \( V_T = V(Y) = K(1 - K) \) (since \( Y \) is a Bernoulli random variable). Let \( G \) be a trait causing locus. For a single locus disease model, the genetic variance due to the locus \( G \) is defined
as \( V_G = V(E(Y|\text{genotype at locus } G)) \). The \textit{Heritability} due to locus \( G \),

\[
0 \leq \gamma = \frac{V_G}{V_T} \leq 1,
\]

is the proportion of variation of the trait attributable to locus \( G \). For a complex disease \( \gamma \) may be interpreted as the locus specific heritability. A discussion about locus specific heritability for binary traits can be found in Lin et al. [2004]. A natural prior for a proportion is the Beta distribution, which is the conjugate prior of the success probability of a Binomial random variable. We assume that

\[
\gamma \sim Beta(\beta_1, \beta_2).
\]

The genetic variance due to a locus can be partitioned into the additive genetic variance, \( V_A \), and the dominance genetic variance, \( V_D \) [Lange, 1997]. Hence, defining the new variables

\[
\alpha = \frac{V_A}{V_T}, \\
\delta = \frac{V_D}{V_T},
\]

we have \( \gamma = \alpha + \delta \). The parameter \( \alpha \) is often referred to as the \textit{Narrow Sense Heritability}. Further, we assume that

\[
\alpha|\gamma \sim Uniform(0, \gamma), \quad (8.2)
\]

that is, given \( \gamma \), we have no preference for values of \( \alpha \), other than that it lies between 0 and \( \gamma \). It also implies that the we do not assume any prior information about the proportion of heritability due to the additive and dominance components; that is, the mode of inheritance of locus \( G \). The probability density function of \( \gamma \) is given by

\[
f(\gamma; \beta_1, \beta_2) = \frac{1}{B(\beta_1, \beta_2)} \gamma^{\beta_1-1}(1 - \gamma)^{\beta_2-1}, \quad (8.3)
\]
where $B(\beta_1, \beta_2)$ is the beta function. The joint distribution of $\alpha$ and $\gamma$, by (8.2), is

$$g(\gamma, \alpha; \beta_1, \beta_2) = \frac{1}{B(\beta_1, \beta_2)} \gamma^{\beta_1-1}(1 - \gamma)^{\beta_2-1} I_{(0,\gamma)}(\alpha),$$  \hspace{1cm} (8.4)

with $I_B(\omega)$ being the indicator function of the set $\{\omega : \omega \in B\}$. The joint prior distribution of $\alpha$ and $\delta$ is obtained from (8.4) by performing the transformation of variables $\delta = \gamma - \alpha$. This transformation leads to the joint distribution of $\alpha$ and $\delta$,

$$h(\alpha, \delta; \beta_1, \beta_2) = \frac{1}{B(\beta_1, \beta_2)} (\alpha + \delta)^{\beta_1-1}(1 - \alpha - \delta)^{\beta_2-1},$$  \hspace{1cm} (8.5)

with $\alpha$, $\delta$ and $\alpha + \delta$ each between 0 and 1.

**Choice of Hyperparameters**

The joint distribution of $\alpha$ and $\delta$ involves the hyperparameters $\beta_1$ and $\beta_2$. It is important to choose values of these two hyperparameters that are reasonable for most complex disease for which our method is proposed. To propose reasonable values of $\beta_1$ and $\beta_2$, we did a survey of the literature to see the estimates of locus specific heritability (a parameter controlled by these hyperparameters) observed for different complex diseases. For Rheumatoid Arthritis (RA), MacGregor et al. [2000] found from twin studies that the heritability of RA is 65% in a Finnish population and 53% in a UK population. Only about one-third of this heritability may be attributable to the HLA locus on chromosome 6 [Deighton et al., 1989]. Such findings put the locus specific heritability of the HLA locus for RA between 17 and 22%. For age-related Maculopathy, Hammond et al. [2002], estimated the heritability of the trait at 45%. If there is one locus that contributes a significant fraction to this heritability, then again we may expect a locus specific heritability of between 10 and 20%. In a study of the heritability of six blood pressure related traits in a founder population, Ober et al. [2001] found heritability values ranging between 21 and 99%. In a heritability anal-
ysis of cytokines as intermediate phenotypes of tuberculosis, Stein et al. [2003] found the intermediate phenotype to have a heritability of 68%. Such findings suggest that the prior distribution of $\gamma$, heritability attributable to a locus, should have a range of $(0, 0.4)$, with peak around 0.1. A choice of $\beta_1 = 2$ and $\beta_2 = 13$ seems to satisfy many of these qualities. In particular, such a distribution has a mode of 0.077, a mean of 0.133, a standard deviation of 0.085 and about 99% of the probability is below 0.4. The choice of these values of $\beta_1$ and $\beta_2$ have another advantage, it leads to a correlation between $\alpha$ and $\delta$ of -0.0845, this agrees well with the genetic reality, given the locus specific heritability, if most of the variability is additive then there is limited scope for dominance variance to contribute, and so on.

### 8.1.4 Sampling Scheme

Let $\pi(l, \alpha, \delta)$ denote the prior distribution of the three parameters. Then, the jump probability for $l$ after $i$ iterations is given by

$$ A(l_i, l'_i; \alpha_i, \delta_i) = \min \left\{ \frac{\pi(l'_i|\alpha_i, \delta_i, G_1, \ldots, G_n)q(l_i'|l_i, \alpha_i, \delta_i)}{\pi(l_i|\alpha_i, \delta_i, G_1, \ldots, G_n)q(l'_i|l'_i, \alpha_i, \delta_i)} \right\} $$

$$ = \min \left\{ 1, \frac{P(G_1, \ldots, G_n|l_i, \alpha_i, \delta_i)\pi(l'_i, \alpha_i, \delta_i)q(l_i'|l_i, \alpha_i, \delta_i)}{P(G_1, \ldots, G_n|l'_i, \alpha_i, \delta_i)\pi(l_i, \alpha_i, \delta_i)q(l_i'|l'_i, \alpha_i, \delta_i)} \right\}, $$

where $l_i$, $\alpha_i$, and $\delta_i$ are the values of the three parameters after $i$ iterations. The proposal distribution, $q(\cdot|l_i, \alpha_i, \delta_i)$, is taken to be uniform on $l_i$ and adjacent locations (three on either side). This leads to symmetric proposals in the interior of the chromosome but not at the tails. The likelihood $P(G_1, \ldots, G_n|l, \alpha, \delta)$ are computed up to a constant using (8.1), and the prior distributions are as defined in the previous section.

The sample from the posterior distribution is obtained using a single component Metropolis algorithm with one complete iteration given by the following steps.
1. Propose a new location $l'$ by sampling from $q(\cdot | l_i, \alpha_i, \delta_i)$. Accept the proposed location with probability $A(l_i, l'; \alpha_i, \delta_i)$.

2. Sample $(\alpha', \delta')$ jointly from a bivariate normal distribution

$$
\begin{pmatrix}
\alpha' \\
\delta'
\end{pmatrix} \sim \mathcal{N}
\begin{pmatrix}
\alpha_i \\
\delta_i
\end{pmatrix},
\begin{pmatrix}
\sigma_0^2 & 0 \\
0 & \sigma_0^2
\end{pmatrix}
$$

Accept $(\alpha', \delta')$ with an acceptance probability similar to the one for $l$.

The parameters $\alpha$ and $\delta$ represent the proportions of trait variance attributable to the additive and dominance components at the locus. As a results, given a fixed heritability, these two parameters are negatively correlated. One way to overcoming possible poor mixing of the Markov chain when parameters in a Bayesian problem are correlated is to jump simultaneously in all parameters [Gilks et al., 1996]. This also speeds up the chain because the likelihood needs to be enumerated a fewer number of times than if the jump was performed in each component separately. Typical values of $\alpha$ and $\delta$ will be in the range $(0, 0.4)$, thus guaranteeing that a value of $\sigma_0 = 0.01$ is sufficient for the chain to mix well. The chain is run for a long time (50,000 iterations are found to be enough), every 50th sample is chosen to get a practically independent sample from the posterior distribution. This sample is used to construct the marginal posterior distributions for $l, \alpha$ and $\delta$. Finally, $100(1 - \epsilon)$ Highest Posterior Density credible intervals are constructed for $l$ [Chen and Shao, 1999].

**Sensitivity Analysis**

Next we explore the effect of the choice of hyperparameters in the posterior distribution of the parameters. Table 8.1 shows the effect of five different choices of $\beta_1$ and $\beta_2$ on the mean and standard deviation (SD) of the marginal posterior distributions. These values are obtained from a MCMC run on 100 or 500 ASPs simulated under the additive disease
model of Table 8.2. As is clear from Table 8.1, the effect of the parameters of the prior
distribution on posterior distribution is limited even with a sample size of 100 ASPs. While
the inference of the parameter $l$ is minimally affected, even $\alpha$ and $\delta$ have small effects due
to $\beta_1$ and $\beta_2$. This suggests that the chosen values of the parameter $\beta_1 = 2$ and $\beta_2 = 13$,
which are motivated by genetic considerations, are suitable for the problem and have a
limited effect on the inference from posterior distributions.

**Convergence of the Chain**

An important part of MCMC methods is to test the convergence of the chain. While for
each chain convergence characteristics will be different, here we try to provide some rules
of thumb to decide *a priori*, how long the chain needs to be run. Figure 8.2 shows the
MCMC sample of $l$, $\alpha$ and $\delta$ for a chain of length 155,500. The chain is constructed on a
sample of size 100 ASPs simulated under the Additive model of Table 8.2. As can be seen
from the Figure, the chain has good mixing properties. We also evaluated the diagnostic
criterion, Potential Scale Reduction Factor (PSRF), proposed by Gelman and Rubin [1992].
Four independent chains with dispersed starting values are run. The MCMC outputs are
thinned (every 50th value taken). Since $\alpha$ and $\delta$ are proportions, a “logit” transform is
applied to convert them to a normal scale. The PSRF point estimates are 1.0, 1.01, and 1.05
for $l$, $\alpha$, and $\delta$, respectively, with 97.5% quantile, given by 1.00, 1.02, and 1.07. Gelman
et al. [2004] suggest that a value of PSRF below 1.1 indicates that the chain has converged.
The suggested criterion is easily satisfied for all the parameters of interest. The analysis of
the convergence properties of the chains suggest that a chain length of 50,000 should be
enough in most situations. Note that, as the sample size increases, the chain could be run
for a smaller number of iterations, since the data will have enough information to guide the
chain to the correct region of the parameter space relatively quickly. However, each chain
needs to be studied individually to see whether it has converged.
8.2 Simulation and Results

**Disease models:** To study the properties of the proposed Bayesian intervals for Linkage Location, we simulate diseases with single and two-locus disease models. Table 8.2 shows three single locus disease models that were simulated. The heritabilities of the three models are comparable and the three models represent Recessive, Additive and Dominant modes of inheritance. The Dominant model is the same as model III in Chapter 6. Models I and II in that Chapter were replaced with two models that have similar modes of inheritance but much lower heritabilities. This was done to use models in the simulations that reflect the modes of inheritance of most complex diseases. In addition to the single locus disease models, the four two locus disease models simulated in Chapter 6 are also simulated. For the three single locus disease models samples of size 100, 250 and 500 ASPs were simulated. For the two-locus disease models 500 ASPs were simulated.

**Confidence set construction tools:** The proposed Bayesian method is compared to four other methods. These are (1) Multi-point CSI-MLS (proposed in Chapter 6); (2) LOD support interval; (3) GEE (LOD support and GEE methods explained in Chapter 4); (4) Bootstrap, the parametric bootstrap method [Papachristou and Lin, 2006c]. Since all these methods are proposed as intermediate fine mapping tools, we shall employ these methods only in regions that have shown evidence of linkage in the coarse map. A chromosome will be assumed to have shown evidence of linkage if the LOD score of Kong and Cox [1997] exceeds the cut-off of 2.33. This threshold is commonly used to flag suggestive linkage.

**Marker maps:** With the advent of cost effective and accurate dense SNP chips, many linkage studies are eliminating the need for microsatellite maps and using dense SNP maps instead. With this development in mind, our simulated chromosomes were assumed to have SNPs every 0.25 cM. Linkage Disequilibrium was not simulated between adjacent SNPs or
between the SNPs and the disease locus. The map was 60 cM long and the disease marker was at 30.125 cM. The step one map, used to provide evidence of linkage in a region and provide estimates of disease parameters for two-step CSI methods, was made up of half the SNPs on the simulated chromosome with a distance between adjacent markers of 0.5 cM. The step two map was made up of the other markers on the chromosome, also with an intermarker distance of 0.5 cM. Each of the two disease chromosomes in the two-locus disease models had the same characteristics. The second map was used by all methods to produce confidence sets. In the last Chapter, we have seen that using SNP maps in both steps of a two-step CSI procedure leads to a small loss of efficiency in the confidence sets thus constructed when compared with microsatellite maps, but the situation reflects most current linkage studies.

**Lengths of nominal confidence sets:** All results are based on 250 replicates. Only those replicates that attained the cut-off of a LOD score of 2.33 in the first map of markers were used with the second map to construct confidence sets. “Coverage” is defined as the ratio of the number of intervals that captured the location of the trait locus, over the number of intervals that signaled linkage on the chromosome (LOD score > 2.33). Tables 8.3 and 8.4 shows the average length of the constructed confidence interval and the observed coverage of the intervals for the single and two-locus disease models. We see that the Bayesian and Bootstrap methods control the coverage to the nominal levels. CSI-MLS has over coverage, while the GEE and LOD-drop methods have under coverage. The Bayesian method leads to the shortest confidence sets for most situations, being smaller than those of GEE and LOD (in most situations), even though the coverage of these two methods is suspect. For the two-locus disease models, LOD performs well in certain situations, with close to correct coverage and precise confidence sets.
**Effect of missing parental genotypes:** Table 8.5 shows the coverage and lengths of nominal 95% confidence sets for the Additive model with 250 ASPs and differing levels of missing parental genotypes (0, 50, 75, and 100%). We see that the coverage probabilities of the Bayesian method and the Bootstrap method are well controlled. The Table shows that missing parental genotypes do not affect the coverage of the Bayesian credible intervals.

**Length of confidence sets with controlled empirical coverage:** Because of the vastly different coverage levels of the different methods, the average lengths of the different methods is not comparable. To be able to compare these lengths, we looked at them when the empirical coverage probability was controlled to the same level for each method. Figures 8.3 and 8.4 shows the average lengths of confidence sets for different levels of empirical coverage. A few observations are clear, the GEE method leads to extremely wide confidence sets in all situations. With small sample sizes, the Bootstrap method does not lead to precise confidence sets. The Bootstrap method also leads to wide confidence sets when the empirical coverage is large. For single locus disease models, the proposed Bayesian method is the best in all situations but the Dominant model, where it is comparable to the LOD-Drop and CSI-MLS. CSI-MLS performs well in most situations, leading to confidence sets only slightly longer than the most superior method. The Bayesian method is inferior to LOD support intervals in four situations: the single locus Dominant model and the second locus of each of EP-2, HET-2 and S-2. Close examination of these disease models shows that the probability that two siblings share 1 allele IBD at the disease locus for each is close to 0.5, which means that $\delta \simeq 0$. Irrespective of this fact, the Bayesian method is estimating it in the model. In these four situations, we performed the Bayesian analysis again by restricting $\delta$ to zero (Bayesian-V2). Table 8.7 shows the results of these new analysis. We see that Bayesian-V2 is better than the original Bayesian methods, but is still inferior to the LOD methods in three situations. While the LOD method is shown to be the most precise in some situations, it is still not clear what drop-size will lead to
a desired level of coverage. Table 8.8 shows the different drop sizes needed to attain an empirical coverage of 0.95. In addition to the disease model and the sample size (as shown in the Table), we believe that the drop size also depends on the marker map used, and the availability of genotypes of related individuals.

**Estimate of disease gene location:** The three methods Bayesian, Bootstrap and GEE also provide location estimates of the disease causing gene. In particular, we use the posterior mode as an estimate of this location in the Bayesian method. Table 8.6 provides the Root Mean Square Errors (RMSE) of the estimate of \( l \) from the three estimation methods. Except for small sample sizes with the Dominant model, the Bayesian estimate is either the best or close to the best of the three methods.

### 8.3 Discussion

We have proposed a Bayesian method of constructing credible intervals for the disease gene location from ASP linkage data. The two-step CSI methods lead to greater than nominal level of coverage and hence may provide confidence sets that are too long to follow up with association and candidate gene methods. The proposed Bayesian method aims to tackle this shortcoming of the CSI methods. The proposed method can use all available genotype data (subject to no Linkage Disequilibrium between markers) and provides confidence sets with nominal levels of coverage. The proposed method has genetically motivated prior distributions, that may be altered based on the experimenter’s prior beliefs with relative ease.

The proposed method is compared to four other methods that have a very similar goal. In the comparisons, the Bayesian method out-performs all the other methods most of the time. Even when it is not the best method, it often provides confidence sets that are close to
that of the most superior method. For diseases with no dominance variance, the LOD-drop approach performs marginally better than the Bayesian method in some situations; however, a vital question that needs to be answered for the LOD support intervals is the size of the drop needed to attain desired coverage. The superior performance of the proposed method to the Bootstrap confidence intervals is also promising. This agrees well with the findings of Manichaikul et al. [2006], who compared the Bootstrap method to an approximate Bayesian method, and observed poor performance of the Bootstrap for the linkage mapping problem. The better performance shows that well motivated prior distributions contribute to efficiency.

Papachristou and Lin [2006c] state that one way to improve the conservative nature of CSI is to integrate out the disease model related parameters instead of using their usually inflated MLEs. By looking at the marginal posterior distribution of \( l \), one is essentially integrating out the disease model related parameters \( \alpha \) and \( \delta \). The observed correct coverage of the proposed method, along with its good performance over a range of disease models, suggests that its use will greatly improve intermediate fine mapping efforts.

The proposed Bayesian method may be extended to include information about covariates into the model. In genetics, covariates often provide information about possible genetic heterogeneity, and use of such available information can greatly help narrow disease gene locations. The parameterization of the likelihood suggested by Olson [1999] could be used to incorporate covariate information into the model.
Hyperparameters $\beta_1 = 2$, $\beta_2 = 13$

Figure 8.1: Plot of the probability density function of $\gamma$ with hyperparameters $\beta_1 = 2$ and $\beta_2 = 13$. 
Figure 8.2: A MCMC run on a sample of 100 ASPs. The three figures show the chain for the three parameters.
Figure 8.3: Comparison of the average lengths of the Bayesian interval, CSI-MLS, GEE, LOD-Drop, and Bootstrap over a range of empirical coverage values.
Table 8.1: Sensitivity of posterior inference to choice of hyperparameters

<table>
<thead>
<tr>
<th># ASP</th>
<th>Prior distribution</th>
<th>Posterior distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β₁, β₂ Mean, SD</td>
<td>l, α, δ Mean, SD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9 0.10 0.09</td>
<td>26.436 3.013 0.091 0.060 0.064 0.042</td>
</tr>
<tr>
<td>2</td>
<td>8 0.20 0.12</td>
<td>25.898 3.514 0.121 0.068 0.062 0.046</td>
</tr>
<tr>
<td>100</td>
<td>2 13 0.13 0.08</td>
<td>26.673 3.112 0.097 0.058 0.092 0.059</td>
</tr>
<tr>
<td>3</td>
<td>12 0.20 0.10</td>
<td>26.128 2.260 0.118 0.061 0.095 0.052</td>
</tr>
<tr>
<td>1</td>
<td>1 0.50 0.29</td>
<td>26.738 2.945 0.127 0.076 0.049 0.037</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9 0.10 0.09</td>
<td>29.954 0.618 0.089 0.026 0.036 0.019</td>
</tr>
<tr>
<td>2</td>
<td>8 0.20 0.12</td>
<td>30.005 0.810 0.080 0.025 0.033 0.018</td>
</tr>
<tr>
<td>500</td>
<td>2 13 0.13 0.08</td>
<td>29.942 0.688 0.076 0.024 0.035 0.018</td>
</tr>
<tr>
<td>3</td>
<td>12 0.20 0.10</td>
<td>29.991 0.801 0.081 0.025 0.035 0.017</td>
</tr>
<tr>
<td>1</td>
<td>1 0.50 0.29</td>
<td>29.985 0.692 0.083 0.025 0.028 0.017</td>
</tr>
</tbody>
</table>

Table 8.2: Single locus disease models and their parameters relevant to the power of the tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>K</th>
<th>H = V₇/V₇, Vₐ/V₇, V₅/V₇, p₅, f₅, f₅, f₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recessive</td>
<td>0.045</td>
<td>0.094 0.014 0.080 0.08 0.04 0.40 0.80</td>
</tr>
<tr>
<td>Additive</td>
<td>0.058</td>
<td>0.110 0.050 0.060 0.10 0.04 0.40 0.80</td>
</tr>
<tr>
<td>Dominant</td>
<td>0.085</td>
<td>0.143 0.143 0.000 0.01 0.07 0.827 0.865</td>
</tr>
</tbody>
</table>

\(^\text{a}\) V₇ is genetic variance of the trait; V₇ is total variance of the trait; H is broad sense heritability; Vₐ and V₅ represent the additive and dominance genetic variance.

\(^\text{b}\) D: High risk allele; d: low risk allele; P₅: allele frequency of D.

\(^\text{c}\) (f₅, f₅, f₅): penetrances of the genotypes (DD, Dd, dd).
Figure 8.4: Comparison of the average lengths of the Bayesian interval, CSI-MLS, GEE, LOD-Drop, and Bootstrap over a range of empirical coverage values for two locus disease models.
Table 8.3: Mean lengths (L) in cM and coverage (CP) of 95% Bayesian credible intervals, CSI-MLS, LOD-drop, GEE, and Bootstrap.

<table>
<thead>
<tr>
<th>Model</th>
<th># ASPs</th>
<th>Bayesian</th>
<th>CSI-MLS</th>
<th>LOD-Drop</th>
<th>GEE</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>L</td>
<td>CP</td>
<td>L</td>
<td>CP</td>
<td>L</td>
</tr>
<tr>
<td>Recessive</td>
<td>100</td>
<td>8.91</td>
<td>0.95</td>
<td>23.03</td>
<td>0.97</td>
<td>11.97</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>3.30</td>
<td>0.96</td>
<td>13.25</td>
<td>1.00</td>
<td>5.17</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.81</td>
<td>0.97</td>
<td>8.69</td>
<td>1.00</td>
<td>2.90</td>
</tr>
<tr>
<td>Additive</td>
<td>100</td>
<td>10.44</td>
<td>0.94</td>
<td>25.57</td>
<td>0.98</td>
<td>11.69</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>6.48</td>
<td>0.94</td>
<td>16.81</td>
<td>1.00</td>
<td>5.90</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2.92</td>
<td>0.95</td>
<td>11.81</td>
<td>1.00</td>
<td>3.32</td>
</tr>
<tr>
<td>Dominant</td>
<td>100</td>
<td>16.10</td>
<td>0.93</td>
<td>26.53</td>
<td>0.99</td>
<td>10.87</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>8.89</td>
<td>0.94</td>
<td>19.55</td>
<td>1.00</td>
<td>6.58</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>4.60</td>
<td>0.93</td>
<td>13.82</td>
<td>1.00</td>
<td>4.05</td>
</tr>
</tbody>
</table>
Table 8.4: Mean lengths (L) in cM and coverage (CP) of 95% Bayesian credible intervals, CSI-MLS, LOD-drop, GEE, and Bootstrap for two locus disease models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Locus</th>
<th>Bayesian L</th>
<th>Bayesian CP</th>
<th>CSI-MLSL</th>
<th>CSI-MLSCP</th>
<th>LOD-Drop L</th>
<th>LOD-Drop CP</th>
<th>GEE L</th>
<th>GEE CP</th>
<th>Bootstrap L</th>
<th>Bootstrap CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP-2</td>
<td>1</td>
<td>3.36</td>
<td>0.94</td>
<td>10.71</td>
<td>1.00</td>
<td>3.61</td>
<td>0.91</td>
<td>5.54</td>
<td>0.85</td>
<td>4.39</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.20</td>
<td>0.94</td>
<td>11.03</td>
<td>1.00</td>
<td>2.36</td>
<td>0.96</td>
<td>5.77</td>
<td>0.85</td>
<td>4.69</td>
<td>0.94</td>
</tr>
<tr>
<td>EP-4</td>
<td>1</td>
<td>10.29</td>
<td>0.95</td>
<td>19.10</td>
<td>0.98</td>
<td>8.83</td>
<td>0.88</td>
<td>9.78</td>
<td>0.84</td>
<td>14.48</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.66</td>
<td>0.96</td>
<td>7.43</td>
<td>1.00</td>
<td>1.88</td>
<td>0.93</td>
<td>3.74</td>
<td>0.87</td>
<td>2.16</td>
<td>0.96</td>
</tr>
<tr>
<td>HET-2</td>
<td>1</td>
<td>3.63</td>
<td>0.94</td>
<td>12.88</td>
<td>1.00</td>
<td>5.09</td>
<td>0.91</td>
<td>6.92</td>
<td>0.88</td>
<td>6.12</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.87</td>
<td>0.93</td>
<td>12.60</td>
<td>1.00</td>
<td>2.85</td>
<td>0.90</td>
<td>6.77</td>
<td>0.86</td>
<td>6.41</td>
<td>0.92</td>
</tr>
<tr>
<td>S-2</td>
<td>1</td>
<td>2.01</td>
<td>0.96</td>
<td>8.71</td>
<td>1.00</td>
<td>3.09</td>
<td>0.95</td>
<td>4.63</td>
<td>0.88</td>
<td>2.79</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.25</td>
<td>0.94</td>
<td>16.79</td>
<td>0.99</td>
<td>4.42</td>
<td>0.92</td>
<td>8.47</td>
<td>0.84</td>
<td>11.36</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Table 8.5: Mean lengths (L) in cM and coverage (CP) of 95% Bayesian credible intervals, CSI-MLS, LOD-drop, GEE, and Bootstrap for the Additive model with 250 ASPs and different levels of missing parental genotypes.

<table>
<thead>
<tr>
<th>Missing (%)</th>
<th>Bayesian</th>
<th>CSI-MLS</th>
<th>LOD-Drop</th>
<th>GEE</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>CP</td>
<td>L</td>
<td>CP</td>
<td>L</td>
</tr>
<tr>
<td>0</td>
<td>6.48</td>
<td>0.94</td>
<td>16.81</td>
<td>1.00</td>
<td>5.90</td>
</tr>
<tr>
<td>50</td>
<td>6.51</td>
<td>0.95</td>
<td>17.35</td>
<td>1.00</td>
<td>7.13</td>
</tr>
<tr>
<td>75</td>
<td>6.76</td>
<td>0.94</td>
<td>18.17</td>
<td>1.00</td>
<td>7.80</td>
</tr>
<tr>
<td>100</td>
<td>7.07</td>
<td>0.96</td>
<td>18.37</td>
<td>1.00</td>
<td>8.36</td>
</tr>
</tbody>
</table>

Table 8.6: Root Mean Squared Error (RMSE) of the estimate of the location (in cM) of the disease gene

<table>
<thead>
<tr>
<th>Model</th>
<th># ASP</th>
<th>RMSE of ( \hat{l} )</th>
<th>GEE</th>
<th>Bayesian</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>3.944</td>
<td>3.077</td>
<td>4.468</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recessive</td>
<td>250</td>
<td>2.514</td>
<td>1.015</td>
<td>1.063</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.730</td>
<td>0.663</td>
<td>0.622</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>4.684</td>
<td>4.333</td>
<td>4.793</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additive</td>
<td>250</td>
<td>3.002</td>
<td>1.900</td>
<td>1.998</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.982</td>
<td>1.160</td>
<td>1.196</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>5.897</td>
<td>6.467</td>
<td>6.196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dominant</td>
<td>250</td>
<td>3.773</td>
<td>3.789</td>
<td>3.099</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2.610</td>
<td>2.083</td>
<td>2.034</td>
<td></td>
</tr>
</tbody>
</table>
Table 8.7: Average lengths of confidence sets with controlled empirical coverage for LOD-drop, Bayesian method and a variation of Bayesian interval, Bayesian-V2 ($\delta$ set to 0)

<table>
<thead>
<tr>
<th>Locus</th>
<th>90%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOD</td>
<td>Bayesian</td>
</tr>
<tr>
<td>Dominant</td>
<td>3.85</td>
<td>3.62</td>
</tr>
<tr>
<td>EP-2 Locus 2</td>
<td>1.98</td>
<td>2.85</td>
</tr>
<tr>
<td>Het-2 Locus 2</td>
<td>2.80</td>
<td>4.04</td>
</tr>
<tr>
<td>S-2 Locus 2</td>
<td>4.17</td>
<td>6.21</td>
</tr>
</tbody>
</table>

Table 8.8: The Drop size needed to obtain a confidence set with 95% empirical coverage tabulated against the disease model and the sample size (# ASPs).

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Recessive</td>
<td>1.39</td>
</tr>
<tr>
<td>Additive</td>
<td>1.17</td>
</tr>
<tr>
<td>Dominant</td>
<td>1.62</td>
</tr>
</tbody>
</table>
Chapter 9

Application of Proposed Methods to Real Data

In this dissertation, we have proposed two novel methods to construct confidence/credible intervals for disease gene locations. Both proposed methods have been shown to have good theoretical properties. We have also studied both methods under a range of simulated models reflecting the complexity of most genetic diseases with unknown etiology. Both methods are also easily applied to real data. In this Chapter, we apply CSI-MLS and the proposed Bayesian method to a real Rheumatoid Arthritis data set.

Rheumatoid Arthritis (RA) is a chronic systemic inflammatory disease of undetermined etiology. The cause is unknown though the presence of auto-antibodies such as anti-cyclic citrullinated peptide (anti-CCP), and an association with human lymphocyte antigen (HLA), have led to its categorization as an autoimmune disorder [Amos et al., 2006]. It has long been suspected that RA has a significant genetic component. MacGregor et al. [2000] used twin studies to establish that the polygenic heritability of RA in UK and Finish populations is 53 and 65%, respectively. Although RA is a complex disease, with possibly multiple loci contributing to the trait [Gregersen, 1999], linkage analysis studies have successfully
mapped only one locus, HLA-DRB1, on chromosome 6 [Mackay et al., 2002, Fisher et al.,
2003, Amos et al., 2006]. The failure to identify additional trait-predisposing loci is be-
lieved to be due to their small contributions, and Amos et al. [2006] have alleviated this
concern to some extent by the use of extremely large sample sizes.

The North American Rheumatoid Arthritis Consortium (NARAC) (http://www.naracdata.org)
is a collaborative effort to understand the genetic basis of RA. Overall, there are 1592 af-
fected siblings contained within 757 multicase RA families. NARAC has performed mi-
crosatellite scans [Jawaheer et al., 2003] using 511 multiplex families. Amos et al. [2006]
retained 642 multicase Caucasian families with 1371 affected siblings, and performed a
SNP linkage scan. Non-Caucasian families were excluded to reduce genetic heterogene-
ity of the study population. A set of 5858 SNPs distributed across the genome is used.
Of these 98.1% passed quality checks resulting in a total of 5744 SNPs. These data
were made available to researchers through the Genetic Analysis Workshop (GAW) 15
(http://www.gaworkshop.org/).

9.1 Data and Methods

Of the possible 757 families, we randomly chose 275 of the Caucasian families, each
with between 4 and 14 individuals. This choice was made in order to have smaller fam-
ilies in the sample, to reflect the situations for which the methods being compared have
been proposed. That lead to 687 affected individuals in the sample and 1325 concor-
dant affected sibling pairs, not all of them independent. Both microsatellite and SNP
data are available on the above families. We use the chromosome 6 microsatellite and
SNP data. A number of confidence set construction methods are compared to see how
precisely each was able to locate the position of HLA-DRB1, a known risk factor for RA.
There are 404 SNPs on the chromosome 6 data. Because marker-marker linkage disequilib-
rium (LD) can lead to biased estimation of IBD sharing probabilities, we used Haploview (http://www.broad.mit.edu/mpg/haploview/) to filter the SNPs to a set in which any two markers had an LD value of $r^2 < 0.02$. This lead to a reduced 320 SNPs on chromosome 6.

Six different methods for construction of confidence sets for the location of the disease gene are compared in the data. They are (1) CSI-Mean: using CSI-V3 of Papachristou and Lin [2006a]; (2) CSI-MLS; (3) The proposed Bayesian method; (4) Bootstrap; (5) LOD drop method; (6) GEE. All of the above methods were used to construct nominal 95% level confidence sets for the disease gene location. A drop size of 1 was used for the LOD-drop approach, since such a drop size is believed to provide a 95% confidence interval. The two CSI methods need two maps, a coarse map and a fine map. Since a microsatellite map was available, we used it as the coarse map to estimate the required disease model related parameters. In particular, we used the MSINT strategy, that is the disease parameter estimates were taken from the location (not just markers) that maximized the LOD score. This strategy has been shown to perform better than other available strategies in Chapter 7.

9.2 Results

Figure 9.1 shows a plot of the LOD score [Kong and Cox, 1997] for a part of chromosome 6. The vertical line denotes the location of HLA-DRB1 locus. As is clear from the figure, the two CSI methods lead to the longest intervals. As also seen in simulation results, CSI-MLS leads to a shorter confidence set than CSI-Mean. GEE has the next longest confidence set, but the observed poor empirical coverage of GEE confidence sets is a caveat against trusting this interval. The Bootstrap and LOD-drop methods provide confidence sets with comparable lengths. The shortest confidence set is provided by the proposed Bayesian method. The chain was explored for convergence and the Potential Scale Reduction Fac-
tors, computed after running four different chains with varied starting points, were 1.01, 1.03, and 1.04, for $l$, $\alpha$, and $\delta$, respectively. The numbers are well below the suggested threshold for concluding convergence.

Figure 9.1: Application of proposed methods to real data. The LOD score plotted for a region of chromosome 6. The different horizontal lines correspond to the different methods used to construct confidence sets for the location of HLA-DRB1.
Appendix A

Confidence Set Inference for the Mean Test

Lin [2002] described the single-point Confidence Set Inference (CSI) approach for two linkage tests statistics, the Mean test and the Two-IBD/Proportion test. An important property for these two statistics to be tested in the CSI framework is for the statistics to be stochastically decreasing in the recombination fraction. While this property is easy to see for the two-IBD test (Lemma B.0.7) it is not trivial for the Mean test statistic. The Mean and Two-IBD tests are defined in Chapter 3. In this Chapter it is assumed that markers are fully informative.

A.1 Stochastic Ordering of Mean Test Statistic

Lin [2002] used the following result.

\[ \sup_{\theta_m \leq \theta_0} P(M_m < C_\alpha | \theta_m) = P(M_m < C_\alpha | \theta_0) \] (A.1)
which can be concluded if $M_n$ is shown to be stochastically decreasing in $\theta$. One way to
guarantee (A.1) is given by Theorem A.1.4, for the proof of which we use the following
definitions and lemmas.

**Definition A.1.1.** Let $X_1$ and $X_2$ be two real-valued random variables, $X_1$ is said to be
stochastically larger than $X_2$ ($X_1 \geq_{st} X_2$) if $F_{X_1}(c) \leq F_{X_2}(c)$, for all $c$.

**Lemma A.1.1.** Let $\{X_1, X_2, \cdots, X_n\}$ be a set of independent random variables and simi-
larly $\{Y_1, Y_2, \cdots, Y_n\}$ be another set of independent random variables. If $X_i \geq_{st} Y_i$, for $i =
1, \cdots, n$, then $\sum_{i=1}^{n} X_i \geq_{st} \sum_{i=1}^{n} Y_i$.  

*Proof.* We provide the proof for $n = 2$, this proof can be extended to the general case. For
any given $c$,

$$
P(X_1 + X_2 \leq c) = \int_{-\infty}^{\infty} P(X_1 \leq c - x_2 | X_2 = x_2) dF_{X_2}(x_2)
$$

$$
= \int_{-\infty}^{\infty} P(X_1 \leq c - x_2) dF_{X_2}(x_2)
$$

$$
\leq \int_{-\infty}^{\infty} P(Y_1 \leq c - x_2) dF_{X_2}(x_2)
$$

Construct a new random variable $X'_2$. Such that $X'_2 =_{st} X_2$ and $X'_2$ is independent of $Y_1$.

This can be easily done by first taking a uniform random variable $U \sim U(0, 1)$ independent
of $Y_1$ and then defining $X'_2 = F_{X_2}^{-1}(U)$.

$$
P(X_1 + X_2 < c) \leq \int_{-\infty}^{\infty} P(Y_1 \leq c - x_2 | X'_2 = x_2) dF_{X'_2}(x_2)
$$

$$
= P(Y_1 + X'_2 \leq c)
$$

$$
= \int_{-\infty}^{\infty} P(X'_2 \leq c - y_1) dF_{Y_1}(y_1)
$$

$$
\leq \int_{-\infty}^{\infty} P(Y_2 \leq c - y_1) dF_{Y_1}(y_1)
$$

$$
= P(Y_1 + Y_2 \leq c).
$$

---

1A special case of theorem 1.A.3 in Shaked and Shantikumar, 1994 (Shaked and Shanthikumar [1994]),
stated without proof.
For the general case note that $X_1 + X_2$ is independent of $X_3$ and so on.

**Definition A.1.2.** A family of CDFs indexed by $\theta$, $F(\cdot; \theta)$, is said to be stochastically increasing (decreasing) in $\theta$ if for $\theta_1 < \theta_2$ we have $F(c; \theta_1) \geq (\leq) F(c; \theta_2)$ for all $c$.

**Lemma A.1.2.** If independent random variables $X_i \sim F_i(\cdot; \theta)$ are stochastically increasing (decreasing) in $\theta$, then the random variable $\sum_{i=1}^n X_i$ is also stochastically increasing (decreasing) in $\theta$.

**Proof.** This follows directly from Lemma A.1.1.

Next define $B_i = \sum_{j=0}^2 j I(mIBD_i = j)$. Then,

$$M_m = \frac{\sum_{i=1}^n B_i}{n}$$

and $E_\theta(B_i) = z_1(\theta) + 2z_2(\theta)$.

$$P_\theta(B_i \leq c) = \begin{cases} 
0 & \text{if } c < 0 \\
z_0(\theta) & \text{if } 0 \leq c < 1 \\
z_0(\theta) + z_1(\theta) & \text{if } 1 \leq c < 2 \\
1 & \text{if } 2 \leq c. 
\end{cases}$$

(A.2)

**Lemma A.1.3.** $z_0(\theta)$ is increasing in $\theta$, $z_2(\theta)$ is decreasing in $\theta$ and $z_0(\theta) + z_1(\theta)$ is increasing in $\theta$.

**Proof.** From [Lin, 2002], we have

$$z_0(\theta) = \frac{K^2 + (1 - \Psi)V_A + (1 - \Psi)^2V_D}{4K^2 + 2V_A + V_D}$$
where $K$ is the population prevalence, $V_A$ the additive genetic variance, $V_D$ the dominant genetic variance and $\Psi = \theta^2 + (1 - \theta)^2$. It is easy to see that since $0 \leq \theta \leq 1/2$,

\[
\frac{d(1-\Psi)}{d\theta} = -2(2\theta - 1) \geq 0
\]
\[
\frac{d(1-\Psi)^2}{d\theta} = -2(1-\Psi)2(2\theta - 1) \geq 0
\]

since $\Psi < 1$ for $\theta > 0$.

Which means that $z_0(\theta)$ is increasing in $\theta$ subject to the condition that $V_A + V_D > 0$. Again from [Lin, 2002] and for $V_A + V_D > 0$, we have

\[
z_2(\theta) = \frac{K^2 + \Psi V_A + \Psi^2 V_D}{4K^2 + 2V_A + V_D}
\]

and

\[
\frac{d\Psi}{d\theta} = 2(2\theta - 1) < 0
\]
\[
\frac{d(1-\Psi)^2}{d\theta} = 2\Psi 2(2\theta - 1) < 0
\]

since $\Psi > 0$.

Which means that $z_2(\theta)$ is decreasing in $\theta$, which is obvious since a larger recombination fraction between the marker and the disease leads to a smaller probability of having two alleles IBD. Next, note that

\[
z_0(\theta) + z_1(\theta) = 1 - z_2(\theta),
\]

meaning $z_0(\theta) + z_1(\theta)$ is increasing in $\theta$. \qed
Theorem A.1.4. $M_m$ is stochastically decreasing in $\theta$, and

$$
\sup_{\theta_m \leq \theta_0} P(M_m < C_\alpha | \theta_m) = P(M_m < C_\alpha | \theta_0).
$$

Proof. By (A.2) and Lemma A.1.3 we see that $B_i$ is stochastically decreasing in $\theta$. By Lemma A.1.2 and because the $B_i$ are independent (assumption of independent ASPs) $\sum B_i$ is stochastically decreasing in $\theta$. It is easy to see that $M_m = \sum B_i / n$ is stochastically decreasing in $\theta$. 

$\square$
Appendix B

Proofs

Lemma B.0.5. The EM algorithm for the MLE of $z$ is given by updates

$$z_i^{(n+1)} = \frac{1}{N} \sum_{j=1}^{N} \frac{z_i^{(n)} w_{ij}}{z_0^{(n)} w_{0j} + z_1^{(n)} w_{1j} + z_2^{(n)} w_{2j}}$$

Proof. Let $Y = (G_1, \ldots, G_N)$ be the incomplete data and $X = (Y, IBD_1, \ldots, IBD_N)$ (where $IBD_i$ denotes the number of alleles shared IBD by the $i^{th}$ ASP) be the complete data. Note that often one may not be able to conclude the number of alleles shared IBD from the marker phenotype data. Then the incomplete data likelihood for the $i^{th}$ ARP is given by $\sum_{j=0}^{2} z_j w_{ji}$. And the complete data log-likelihood is given by

$$\ell = \sum_{i=1}^{N} \ln P_z(G_i, IBD_i)$$

$$= \sum_{i=1}^{N} \ln (P_z(G_i|IBD_i)P(IBD_i))$$

$$= \sum_{i=1}^{N} \ln \prod_{j=0}^{2} (w_{ji}z_j)^{I(IBD_i)}$$

$$= \sum_{i=1}^{N} \sum_{j=0}^{2} I(IBD_j) \ln(w_{ji}z_j)$$
E-Step: At the $n^{th}$ iteration in the E-step we shall calculate the conditional expectation of 
(the notation used is consistent with Chapter 2 of Lange [1997])

$$Q(z|z^{(n)}) = E(\ln L(z; X)|Y, z^{(n)}),$$

which reduces to evaluating

$$E(I(IBD_i = j)|G_i, z^{(n)}) = \frac{P(G_i|IBD_i = j, z^{(n)}) P(IBD_i = j|z^{(n)})}{P(G_i|z^{(n)})} = \frac{w_{ji}z_{j}^{(n)}}{\sum_{j=0}^{2} w_{ji}z_{j}^{(n)}}.$$

Therefore we have,

$$Q(z|z^{(n)}) = \sum_{i=1}^{N} \sum_{j=0}^{2} \frac{w_{ji}z_{j}^{(n)}}{\sum_{j=0}^{2} w_{ji}z_{j}^{(n)}} \ln (w_{ji}z_{i}).$$

M-Step: Maximize $Q(z|z^{(n)})$ with respect to $z$. Define

$$k_{ji}^{(n)} = \frac{w_{ji}z_{j}^{(n)}}{\sum_{j=0}^{2} w_{ji}z_{j}^{(n)}}.$$

We have

$$\frac{\partial Q(z|z^{(n)})}{\partial z_1} = \frac{\partial}{\partial z_1} \sum_{i=1}^{N} \sum_{j=0}^{2} k_{ji} \ln (w_{ji}z_{i})$$

$$= \sum_{i=1}^{N} \left\{ k_{1i}^{(n)} \frac{\partial \ln z_1}{\partial z_1} + k_{2i}^{(n)} \frac{\partial \ln (1 - z_1 - z_2)}{\partial z_1} \right\}$$

$$= \sum_{i=1}^{N} \left\{ \frac{k_{1i}^{(n)}}{z_1} - \frac{k_{2i}^{(n)}}{1 - z_1 - z_2} \right\}$$

$$= 0.$$
From the similar derivative with respect to $z_2$ and because $a/b = c/d \Rightarrow a/b = (a+c)/(b+d)$ we have

$$
\sum_{i=1}^{N} k_{1i}^{(n)} z_1 = \sum_{i=1}^{N} k_{2i}^{(n)} z_2 = \sum_{i=1}^{N} k_{0i}^{(n)} (1 - z_1 - z_2) = \sum_{j=0}^{2} \sum_{i=1}^{N} k_{1i}^{(n)} z_1 - \sum_{i=1}^{N} k_{1i}^{(n)} = n,
$$

which proves the lemma.

**Lemma B.0.6.** $z_2(1/2) = 1/4$ and $z_2(\theta)$ is decreasing for $\theta \in (0, 1/2]$.

**Proof.** When the marker is unlinked to the disease the IBD sharing probabilities follow the null, meaning $z_2(1/2) = 1/4$. To show the next part note that,

$$
\Psi = \theta^2 + (1 - \theta)^2
$$

$$
\frac{d\Psi}{d\theta} = 2\theta - 2(1 - \theta)
$$

$$
= 2(2\theta - 1).
$$

This implies,

$$
\frac{dz_2(\theta)}{d\theta} = \frac{1}{4K^2 + 2V_A + V_D} \frac{d}{d\theta}(K^2 + \Psi V_A + \Psi^2 V_D)
$$

$$
= \frac{1}{4K^2 + 2V_A + V_D}(4V_D \Psi(2\theta - 1) + 2V_A(2\theta - 1)).
$$

Note that for $\theta \in (0, 1/2)$ and for the condition $V_A + V_D > 0$ we have $\frac{dz_2(\theta)}{d\theta} < 0$. Hence $z_2(\theta)$ is decreasing for $\theta \in (0, 1/2]$.

**Lemma B.0.7.** Let $X \sim Binomial(n, p)$, if $p_1 < p_2$ then we have $P_{p_1}(X < c) \geq P_{p_2}(X < c)$, for all $c \geq 0$.

**Proof.** Note that

$$
P_{p_1}(X < c) \geq P_{p_2}(X < c) \Leftrightarrow P_{p_1}(X \geq c) \leq P_{p_2}(X \geq c).
$$
From the relation between the incomplete Binomial sum and the incomplete Beta integral [Weisstein, 2006].

\[ B(x; a, b) = \int_0^x u^{a-1}(1 - u)^{b-1}du \] and \( B(1; a, b) = B(a, b) \).

\[
P_{p_1}(X \geq c) = \sum_{x=[c]}^{n} \binom{n}{x} p_1^x (1 - p_1)^{n-x}
\]

\[
= \frac{B(p_1; [c], n - [c] - 1)}{B([c], n - [c] - 1)}
\]

\[
= \frac{\int_0^{p_1} u^{[c]-1}(1 - u)^{n-[c]-2}du}{\int_0^{p_2} u^{[c]-1}(1 - u)^{n-[c]-2}du}
\]

\[
< \frac{B([c], n - [c] - 1)}{B([c], n - [c] - 1)} \quad \text{(Since the integrand is positive.)}
\]

\[
= P_{p_2}(X \geq c).
\]

\[ \square \]

**Lemma B.0.8.** The parametric estimates of \( z_i(\theta) \) obtained from equation (6.4) satisfy the possible triangle constraints

**Proof.** By the relation between the relative risks, \( \lambda_S \) and \( \lambda_O \), and \( K \) (population prevalence), \( V_A \) (additive genetic variance) and \( V_D \) (dominant genetic variance) [James, 1971] we can represent \( z_i(\theta) \) as,

\[
\begin{align*}
z_0(\theta) &= \frac{K^2 + (1 - \Psi)V_A + (1 - \Psi)^2V_D}{4K^2 + 2V_A + V_D}, \\
z_1(\theta) &= \frac{2K^2 + V_A + 2\Psi(1 - \Psi)V_D}{4K^2 + 2V_A + V_D}, \\
z_2(\theta) &= \frac{K^2 + \Psi V_A + \Psi^2 V_D}{4K^2 + 2V_A + V_D}.
\end{align*}
\]

Using the relations above we need to show that the possible triangle constraints [Holmans,
1993] are satisfied, in other words, the following three inequalities are satisfied.

\[ z_1(\theta) \geq 2z_0(\theta) \quad \text{(B.2)} \]
\[ z_1(\theta) \leq \frac{1}{2} \quad \text{(B.3)} \]
\[ z_0(\theta) \geq 0. \quad \text{(B.4)} \]

First note that

\[ \frac{d\Psi}{d\theta} \quad = \quad 2(2\theta - 1) < 0 \quad \text{for } 0 < \theta < \frac{1}{2} \]
\[ \frac{d\Psi(1 - \Psi)}{d\theta} \quad = \quad 2(2\theta - 1)(1 - 2(\theta^2 + (1 - \theta)^2)) > 0 \quad \text{for } 0 < \theta < \frac{1}{2} \]

Note that \( \Psi(0) = 1 \) and \( \Psi(1/2) = 1/2 \), also \( \Psi \) is decreasing in \((0, 1/2)\). Which means that \((1 - \Psi) \geq 0 \) in \((0, 1/2)\). Thus we have

\[ z_0(\theta) = \frac{K^2 + (1 - \Psi)V_A + (1 - \Psi)^2V_D}{4K^2 + 2V_A + V_D} \geq 0 \quad \text{for } 0 < \theta < \frac{1}{2}. \]

Further, \( \Psi(1 - \Psi) \) is increasing in \((0, 1/2)\) and \( \Psi(1/2)(1 - \Psi(1/2)) = 1/4 \), hence

\[ z_1(\theta) = \frac{2K^2 + V_A + 2\Psi(1 - \Psi)V_D}{4K^2 + 2V_A + V_D} \]
\[ \leq \frac{2K^2 + V_A + \frac{1}{2}V_D}{4K^2 + 2V_A + V_D} \]
\[ = \frac{1}{2} \quad \text{for } 0 < \theta < \frac{1}{2} \]

Also it can be seen that \( 2\Psi \geq 1 \) and \( \Psi \leq 1 \), which guarantees that

\[ z_1(\theta) - 2z_0(\theta) = \frac{V_A(2\Psi - 1) + V_D2(1 - \Psi)(2\Psi - 1)}{4K^2 + 2V_A + V_D} \geq 0 \quad \text{for } 0 < \theta < \frac{1}{2}. \]
Bibliography


