Genetic linkage analysis based on identity by descent using Markov chain Monte Carlo sampling on large pedigrees

by

Alexandre Bureau

B.S. (Université Laval) 1996
M.A. (University of California, Berkeley) 1998

A dissertation submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy

in

Biostatistics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Terence P. Speed, Chair
Professor Mark van der Laan
Professor Montgomery Slatkin

Spring 2001
The dissertation of Alexandre Bureau is approved:

Chair

Date

Date

Date

University of California, Berkeley

Spring 2001
Genetic linkage analysis based on identity by descent using Markov chain Monte Carlo sampling on large pedigrees

Copyright 2001

by

Alexandre Bureau
Abstract

Genetic linkage analysis based on identity by descent using Markov chain Monte Carlo sampling on large pedigrees

by

Alexandre Bureau
Doctor of Philosophy in Biostatistics
University of California, Berkeley
Professor Terence P. Speed, Chair

In large pedigrees, the informativity of genetic markers for linkage analysis is limited by missing data on ancestors, and methods exploiting the increased identity by descent (IBD) sharing between affected individuals at a disease locus are not as well understood as in small pedigrees. We infer gene transmission from multiple linked markers to increase the power of linkage analysis on large pedigrees. Multilocus inheritance is represented by a hidden Markov model where the observed data are marker phenotypes and the hidden states are vectors of meiosis indicators. Since exact multilocus analysis is not computationally feasible on large pedigrees, a Markov chain Monte Carlo (MCMC) sampling approach is developed to obtain approximations, involving the meiosis and locus Gibbs samplers and new Metropolis samplers. Simulations show that combining multiple samplers improves the precision and accuracy of MCMC estimates. We also consider the choice of statistics measuring IBD sharing. Focusing on the linkage analysis of glaucoma in a large pedigree from Tasmania, we design statistics based on the hypotheses that a disease allele is shared by a large number of affecteds either across the entire pedigree or in two sub-pedigrees. Those statistics are found to be powerful with complete IBD information, but no more than the generic statistic $S_{pairs}$. At a locus where many affected individuals share IBD, simulations reveal that the approximation of IBD sharing statistics by their conditional expectation given marker data improves as the number of linked marker loci utilized increases, but the conditional expectation remains far below the value of the statistic. The analysis of a genome scan on the 246 member glaucoma pedigree revealed that the MCMC samplers mix too slowly.
in a problem of that scale to give reliable estimates within a reasonable computing time. The MCMC samplers nonetheless inferred correctly a subset of affected individuals sharing IBD near a locus where they inherited an ancestral mutation. The analysis of the genome scan points to a region where a large number of affecteds may share IBD, but divergent estimates from multiple runs prevent us from reaching a definite conclusion on the actual IBD sharing level.

Professor Terence P. Speed  
Dissertation Committee Chair
To Sylvie, my Angel.

À mes tendres parents, Hélène et Réjean.
# Contents

List of Figures vii  
List of Tables x

1 Introduction 1  
1.1 Genetics background 2  
1.1.1 Notation 6  
1.2 Outline 6

2 Genetic linkage analysis with pedigrees 9  
2.1 The meiosis indicator representation of inheritance 9  
2.2 Identity by descent sharing and its role in linkage analysis 11  
2.2.1 Illustration of the increased probability of IBD sharing between affected relatives 12  
2.2.2 Identity by descent configuration 14  
2.2.3 Review of allele sharing methods of linkage analysis in general pedigrees 16  
2.3 Assessing significance of identity-by-descent sharing test statistics 19  
2.3.1 Genomewide significance 19  
2.4 Missing inheritance information in pedigrees 20  
2.4.1 Inferring identity by descent 21  
2.5 A hidden Markov model of multilocus inheritance 22  
2.5.1 Inferring conditional distribution of meiosis indicators 25  
2.6 IBD sharing statistics under incomplete information 26

3 Markov chain Monte Carlo sampling of meiosis indicators 30  
3.1 Whole meiosis Gibbs sampler 33  
3.2 Non-communicating classes of states in nuclear families with the meiosis Gibbs sampler 34  
3.3 Updating all meioses between a parent and his children 39  
3.4 Whole locus Gibbs sampler 40  
3.4.1 Pedigree peeling 41  
3.4.2 Pedigree peeling at one locus conditional on meiosis indicators at flanking loci 43  
3.4.3 Sampling of the ordered genotypes 45
3.4.4 Sampling of the meiosis indicators ........................................ 46
3.5 Generating an initial configuration of meiosis indicators ............... 46
3.6 Reinitializing the chain in a Metropolis step .............................. 47
3.7 Monitoring convergence of estimates ....................................... 49
  3.7.1 Convergence diagnostic tools ........................................ 51
3.8 Evaluation of convergence of different sampler combinations ......... 56

4 Searching for a gene predisposing to glaucoma in a pedigree from Tasmania
  4.1 Primary open-angle glaucoma ............................................. 67
  4.2 The pedigree GTAS2 ........................................................ 68
    4.2.1 Available marker data ............................................. 70
    4.2.2 Hypotheses on the genetic causes of glaucoma in the affected individuals of the pedigree GTAS2 ............... 72
  4.3 Largest set of alleles IBD in affected individuals ...................... 76
    4.3.1 Extension to potentially heterogeneous sub-pedigrees ............ 79
  4.4 Power of IBD sharing statistics ........................................ 81
    4.4.1 Power of IBD sharing statistics under the hypotheses on genetic causes of glaucoma in the pedigree GTAS2 ............... 82
    4.4.2 Estimation of IBD sharing statistics by their conditional expectation computed in a multilocus analysis using different numbers of markers ............... 84
  4.5 Parameters of the genome scan analysis ................................ 90
  4.6 Convergence monitoring for the analysis of the genome scan on the pedigree GTAS2 .......................................................... 91
    4.6.1 Statistics monitored ................................................ 91
    4.6.2 Mixing diagnostics results ......................................... 91
    4.6.3 Standard deviation estimates ..................................... 94
    4.6.4 Relationship between variance of the estimates and mixing .... 98
    4.6.5 Influence of marker information on variance of estimates and mixing speed ................................................. 98
    4.6.6 Convergence at the level of condensed IBD configurations .... 102
    4.6.7 Variability between runs ........................................... 103
    4.6.8 Conclusion of the convergence monitoring ........................ 105
  4.7 Inference on IBD sharing at the GLC1A locus ............................ 107
  4.8 Results of the IBD sharing analysis of the genome scan ................ 109
    4.8.1 Comparison of the distribution of IBD sharing statistics estimates at marker loci of the genome scan to the null distribution of the statistics ............................. 110
    4.8.2 Follow-up of the region with the highest signal .................. 112

5 Discussion and conclusion ...................................................... 120

Bibliography ................................................................. 126

A Computing the probability of marker data given the meiosis indicators at a locus 132
B  MCMC sampling software  137
## List of Figures

1.1 Schematic illustration of a crossover between a pair of homologous chromosomes during the prophase of meiosis. ................................. 4  
1.2 Marker phenotypes and meiosis indicators at three loci. .............. 10  
1.3 Example of IBD configuration between two sibs and an aunt. ........... 15  
1.4 Graphical representation of the hidden Markov model formed by the marker phenotype (filled circles) and meiosis indicator vectors (empty circles) at a sequence of linked loci. ................................. 25  

3.1 Two sibs with different alleles. ........................................... 36  
3.2 Three sibs with genotypes such that each child has one allele in common with each of his two siblings. .............................................. 37  
3.3 Four sibs with genotypes such that each child has one allele in common with two of his siblings and none with his third sibling. ................. 38  
3.4 Diagram of gene transmission from parent to offspring. ............... 44  
3.5 Pedigree and marker phenotypes for test case 1. .......................... 58  
3.6 Pedigree used for test case 2. ............................................. 59  
3.7 Cusum plot of $\hat{Z}$ with $Z = S_{pairs}$ for different proportions of locus and meiosis samplers in test case 1 at locus 5. ......................... 62  
3.8 Cusum plot of $\hat{Z}$ with $Z = S_{pairs}$ for different proportions of locus, meiosis and parental samplers in test case 1 at locus 5. ......... 63  

3.9 Cusum plot of $\hat{Z}$ with $Z = S_{pairs}$ for the locus, locus + meiosis, locus + meiosis + parental and locus + chain restart + meiosis + parental samplers in test case 1 at locus 5. ............................................. 64  
3.10 Cusum plot of likelihood ratio for the locus, locus + meiosis, locus + meiosis + parental and locus + chain restart + meiosis + parental samplers in test case 2 at location 18cM. ............................................. 65  

4.1 Pedigree GTAS2. .................................................................. 71  
4.2 Genetic map of the markers used in the genome scan of the GIST. .... 75  
4.3 Number of allele types seen in the phenotypes of individuals from the pedigree GTAS2 at 401 marker loci. ....................................... 75  
4.4 Power of IBD sharing statistics under scenarios 1 and 2 on pedigree GTAS2. 83  
4.5 Power of IBD sharing statistics under scenarios 3a and b on pedigree GTAS2. 84
4.6 Genetic map of the region of chromosome 10 used for data simulation. . . .
4.7 Mean of the actual values and conditional expectations of $S_{most}$ in 20 replicates of the phenotype data on the pedigree of figure 3.6.  
4.8 Mean of the actual values and conditional expectations of $S_{pairs}$ in 20 replicates of the phenotype data on the pedigree of figure 3.6. 
4.9 Estimated mean difference between $\bar{S}_{most}$ computed using six and three markers and difference between six and one marker at six loci on chromosome 10 for the pedigree of figure 3.6 with simultaneous 95% confidence intervals for the two differences at each locus separately. 
4.10 Estimated mean difference between $\bar{S}_{most}$ computed using six and three markers and difference between six and one marker at six loci on chromosome 10 for the pedigree of figure 3.6 with simultaneous 95% confidence intervals for the two differences at each locus separately. 
4.11 Statistic D for $S_{most}$ and $S_{pairs}$ for the set of 24 affected individuals in the pedigree GTAS2. 
4.12 Cusum plot for $S_{most}$ at D6S422. 
4.13 Cusum plot for $S_{pairs}$ at D6S422. 
4.14 Cusum plot for $S_{most}$ at D2S396. 
4.15 Cusum plot for $S_{pairs}$ at D2S396. 
4.16 Log base 10 of the ratio of the batch means estimator of the SD of $\hat{S}_{pairs}$ over the window estimator against the window estimator. 
4.17 Log base 10 of the ratio of the batch means estimator of the SD of $\hat{S}_{most}$ over the window estimator against the window estimator. 
4.18 Window estimate of the SD of $\hat{S}_{most}$ against $D$ statistic. 
4.19 Window estimate of the SD of $\hat{S}_{pairs}$ against $D$ statistic. 
4.20 Window estimate of the SD of $\hat{S}_{most}$ against number of observed heterozygotes. 
4.21 Window estimate of the SD of $\hat{S}_{pairs}$ against number of observed heterozygotes. 
4.22 $D$ statistic for $\hat{S}_{most}$ against number of observed heterozygotes. 
4.23 $D$ statistic for $\hat{S}_{pairs}$ against number of observed heterozygotes. 
4.24 Estimate of the conditional expectation of the statistic $S_{most}$ on the 24 affected individuals of the pedigree GTAS2 for chromosome 6 in two MCMC runs. 
4.25 $\chi^2$ distance between the marginal distributions of condensed IBD configurations at every marker locus on chromosome 6 in two MCMC runs. 
4.26 Estimates of the conditional expectation of $S_{pairs}$ for 24 affecteds in the pedigree GTAS2. 
4.27 Estimates of the conditional expectation of $S_{most}$ for 24 affecteds in the pedigree GTAS2. 
4.28 Estimates of the conditional expectation of $S_{pairs}$ for 15 affecteds in the pedigree GTAS2. 
4.29 Estimates of the conditional expectation of $S_{most}$ for 15 affecteds in the pedigree GTAS2. 
4.30 Estimates of the conditional expectation of the sum of $S_{most}$ in two subpedigrees of GTAS2 with 19 and 5 affected individuals.
4.31 empirical distribution of $S_{most}$ at 401 marker loci against null distribution of $S_{most}$ for 24 affecteds in the pedigree GTAS2. ................................. 116
4.32 empirical distribution of $S_{pairs}$ at 401 marker loci against null distribution of $S_{pairs}$ for 24 affecteds in the pedigree GTAS2. ................................. 116
4.33 Estimates of the conditional expectation of $S_{most}$ for 24 affecteds at four markers on chromosome 3 from several MCMC runs. ................. 117
4.34 Cusum plot of the first 18,000 consecutive realizations of $S_{most}$ on 24 affecteds at locus $M_4$. ................................................................. 118
4.35 Sampled values of $S_{most}$ for 24 affecteds at $M_4$ from two MCMC runs. ................................. 118
4.36 Distribution of $\log(P[S|Y])$ for the sampled realizations of $S$ for four marker loci on chromosome 3 in two MCMC runs, one started from an initial state leading to high values of IBD sharing statistics (high sharing) and the other from an initial state leading to values of IBD sharing statistics near their expectation (low sharing). ................................. 119

A.1 Pedigree and marker data used in the example of computation of the probability of marker data given the meiosis indicators at a locus. .................... 133
A.2 Examples of descent graph and its corresponding founder graph. ............ 134
## List of Tables

3.1 Joint genotypes of a sibship such that the genotypes of the sibs are all distinct and there are 4 possible individual genotypes. The numbers refer to the state sets described in the text. “Other” refers to a state set that is not described because it is irreducible with single meiosis flips. .................................................. 36

3.2 Meiosis vectors of the two sibs on figure 3.1 consistent with their observed genotypes. .................................................................................................................. 37

3.3 Meiosis vectors of the four sibs on figure 3.3 consistent with their observed genotypes. .................................................................................................................. 39

3.4 Joint probabilities of meiosis indicators and allele identity events at one locus. 44

3.5 Convergence of $Z$ with $Z = S_{pairs}$ for different combinations of MCMC samplers in test case 1. L = locus, M = meiosis, P = parental, R = chain restart. .................................................................................................................. 61

3.6 Convergence of likelihood ratio with different combinations of MCMC samplers in test case 2. L = locus, M = meiosis, P = parental, R = chain restart. .................................................................................................................. 61

4.1 Penetrances of the quasi-dominant one-gene and two-gene models for the power study on the pedigree GTAS2. ................................................................. 82

4.2 Penetrances used to simulate disease phenotypes. ................................................. 85

4.3 Median and maximum estimates of the SD of $S_{most}$ and $S_{pairs}$ at the genome scan loci. .................................................................................................................. 96

4.4 Counts of the 5 most frequent condensed IBD configurations at locus D6S422.103

4.5 Counts of the 5 most frequent condensed IBD configurations at locus D2S396.103

4.6 Highest probability largest subset of affecteds sharing an allele IBD for each of 5 loci around the GLC1A locus conditional on data at 34 marker loci on chromosome 1 in a first MCMC run. ............................................................... 108

4.7 Highest probability largest subset of affecteds sharing an allele IBD for each of 5 loci around the GLC1A locus conditional on data at 34 marker loci on chromosome 1 in a second MCMC run. ......................................................... 109

A.1 Compatible allele assignments for the components of the example founder graph. .................................................................................................................. 136
Acknowledgements

The completion of this work was a great challenge for me. I owe many remarkable people for helping and supporting me accomplish my goal. My thanks go first to Terry Speed who communicated to me his inexhaustible energy and enthusiasm for the fascinating field of statistical genetics. Terry, you never failed to boost my morale when I doubted my capacity to finish my Ph.D. and you pushed me to maintain scientific rigor when I was being careless. The trip to Australia to visit the Walter and Eliza Hall Institute was a wonderful bonus to the privilege of working with you.

Monty Slatkin’s lectures were a captivating introduction to genetic epidemiology and human population genetics concepts in my first year, and I am grateful to him for accepting to serve on my dissertation committee. From Mark van der Laan, I learned much about the fundamental theoretical principles of statistical estimation in his survival analysis class. I appreciated his friendliness has a teacher and an advisor. My stay at Berkeley was enriched by the teaching of many other dedicated professors. David Brillinger’s lectures were just as lively as his conversations in the lunch room, where he would always have a Canadian anecdote to share with me. Others that I want to thank for communicating their knowledge to me include Peter Bickel, Richard Brand, David Giltinan, Allan Hubbard, Michael Jordan, Maureen Lahiff, Dan Moore and John Rice.

I am indebted to Paul Baird for involving me in the genetic study of the pedigrees of the Glaucoma Inheritance Study in Tasmania and providing me the data of the genome scan analyzed in chapter 4. I learned from his patient explanations the inner workings of a genetic study while I was at WEHI. I must also direct my thanks to Steve Shiboski, who supervised my work as a graduate student researcher at UCSF. He introduced me to scientific research and the arduous task of writing papers for scientific journals.

Computing was at the heart of my research, and the support of the Statistical Computing Facility staff has been essential to the conduct of my work. Special thanks to Phil Spector for fulfilling every software request that I made over the years amazingly quickly, sometimes in a matter of minutes. The handling of administrative matters related to my graduate studies was made much easier by the diligent work of Bonnie Hutchings,
our graduate assistant. I appreciated her availability and responsiveness to my numerous requests since the day I was accepted into the program. Sara Wong and Angie Fong from the Department of Statistics were also of great help.

My student life was made more enjoyable by the company of my fellow graduate students. Although I found my closest friends in Berkeley outside of school, I am glad to have known and shared my work days with many students in the statistics and biostatistics programs. Sandrine Dudoit was my guide as I was delving into the field of linkage analysis, and her work is the model of scientific insight and thoroughness that I could only dream to achieve. It was always a pleasure chatting with my fellow Québécoise in the department. I enjoyed the spirit of Simon Cawley; we had fun when we were together in Melbourne. I shared good moments with Von Bing Yap, Nicola Armstrong, Yee Wha Yang and Ingileif Hallgrimsdottir. I want to thank Ingileif specially for reviewing parts of this manuscript.

In the course of five years in the same office, numerous office mates came and went. My best memories are of the time when Kristin Mork and Michael Moser occupied the other two desks and spontaneous conversations would erupt at any time. I have somewhat drifted away from the biostatistics students in the last few years, but I have not forgotten the good times I had during my masters studies with Maja Pavlić, Jenny Brian and Tu Duong.

My academic learning was coupled with learning about life. The foremost life experience of those years started as a timid flirt to grow into an ever deepening and fulfilling relationship with my dear love Sylvie Marceau. Since our first Halloween date she has been the companion of my adventures, from the frozen expanses of Grand-Falls to the colorful splendor of the Great Barrier Reef. I would not have gone through the periods of uncertainty and discouragement without her comforting presence. Sylvie, the thought that you will become my wife fills me with intense happiness.

I came to Berkeley straight from the warmth of the home of my cherished parents Hélène et Réjean. Their love and support has always been with me and despite the tears of separation, they encouraged me to pursue my dreams in California. Maman et Papa, je vous aime et vous dois beaucoup de mon succès et mon bonheur. La pensée du nid douillet que vous me gardez m’aide à voler de mes propres ailes.
Chapter 1

Introduction

The combination of genes that we inherited from our parents contributes to our physical and psychological traits. Defects in one or more genes may cause or predispose to diseases. An important step in understanding the biology of genetic diseases is to identify the responsible genes. The benefits from that knowledge are multiple. Biochemical assays can be designed to determine the variants of the genes carried by an individual and that information used to evaluate the risk of developing a disease like cancer, heart disease or Alzheimer disease and target prevention efforts toward susceptible individuals. It can also help diagnosis on people manifesting symptoms and prevent the birth of children with a lethal condition. Knowledge of the genes involved in a disease leads to the identification of missing or over-abundant proteins and the development of drugs to replace or block them. Ultimately, it may be possible to develop a gene therapy by which normal copies of a gene are inserted into the genome of patients carrier of a defective gene variant.

The association between disease occurrence and transmission of the genetic material in families is the basis for linkage analysis, a method to locate the position of disease genes. In this thesis we address computational and statistical issues arising when linkage analysis is applied to large multi-generation pedigrees, with a particular focus on the analysis of such a pedigree from a study of glaucoma.

This introductory chapter presents basic concepts of genetics. In particular, we explain the phenomenon of genetic linkage and the concept of genetic distance. We then give an outline of the thesis.
1.1 Genetics background

The information to produce a human being is contained in molecules of deoxyribo-
nucleic acid (DNA). The genetic information is encoded in the linear sequence of four
types of bases composing a DNA strand: adenine (A), guanine (G), cytosine (C) and
thymine (T). Long molecules of double-stranded DNA in the nucleus of the cell are called
chromosomes. Humans have 23 pairs of chromosomes, 22 homologous pairs of autosomes
and a pair of sex chromosomes, a X and a Y in males and two X’s in females. Each indi-
vidual inherits a set of 23 chromosomes containing a copy of the human genome from each
parent.

Genes are segments of DNA sequence translated into proteins. Proteins are the
molecules forming the structure of living organisms and performing most of the activities
of life, from transforming energy into movement to replicating DNA. Genes represent an
estimated 1.5% of the DNA sequence in the human genome. The entire sequence of the
human genome is now deciphered and the genes are in the process of being identified. It is
now estimated that the human genome contains about 30,000 genes (International Human
Genome Sequencing Consortium [5]).

Through the action of the protein it encodes, a gene influences observable traits of
an organism. Knowing the sequence of a gene provides clues about the biological function
of the coded protein, but it does not reveal its final effect on the traits of the organism. Geneticists study that effect by relating trait variability between individuals to variants in
the DNA sequence of genes and their regulatory regions. The observed value of a trait
exhibiting variability due to differences in gene sequences is called a phenotype. Of primary
interest in health research are pathological manifestations. A phenotype may either be cat-
egorical, like being affected or unaffected by a disease, or quantitative, like blood cholesterol
level.

A locus (plural loci) is a position on the genome treated as a point at the scale of
entire chromosomes, but in fact consisting of a segment of DNA. DNA sequence variants at
a locus are called alleles, a term also applied by extension to the DNA copies themselves.
For a particular locus, the genotype of an individual consists of two copies of the DNA
segment, one on each homologous chromosome in a pair, whose sequence corresponds to
one of the allele types. A genotype consisting of two identical alleles is homozygous while
a genotype made of two different alleles is heterozygous. The genotypes at loci on the sex
chromosomes in male are exception to the rule, consisting of a single DNA segment on the
only copy of the X and Y chromosomes.

The conditional probability of a phenotype given the genotype at one or multiple
loci is called penetrance. A genetic model for a trait is a specification of the penetrances of
the genotypes of one or more genes having an effect on the trait. Simple traits are governed
by a single gene, and the genotype of that gene determines or at least strongly predisposes
to the phenotype. A phenotype that is manifested with high probability when either one or
two copies of a particular allele are present is called dominant and the allele type causing
it is said to have a dominant effect. The term recessive applies to phenotypes expressed
only when two copies of the same allele are present. In either case the penetrance of a
 genotype may be complete, i.e. equal to one, or incomplete. Complex traits are determined
by complex interactions between multiple genes and also environmental factors. The effect
of an allele of a gene is then only to modify the probability of manifesting a phenotype, and
the marginal penetrances of single gene genotypes are all non-zero. A genetic model for one
of the genes involved is quasi-dominant if the presence at the gene locus of one copy of an
allele associated to the phenotype modifies the penetrance and it is quasi-recessive if two
copies of the allele are needed to change the penetrance compared to the baseline.

A positive association tends to be observed between the phenotypes of related in-
dividuals because some of their genes are copies of ancestral genes. Such genes are said to be
identical by descent (IBD). With our knowledge of the mechanisms of genetic inheritance,
that association can be exploited to find the genomic location of genes influencing a trait
without prior information on the genomic DNA sequence. Genetic linkage analysis is the
method to locate genes based on the dependence, or linkage, between the alleles transmitted
from a parent to his offsprings for genes near each other on a chromosome.

Meiosis is the process of cell division producing the gametes (egg or sperm). Dur-
ing meiosis, the paternal and maternal copies of a chromosome exchange DNA, by breaking
and rejoining in what are called crossovers, and a mosaic of the two parental chromosomes
is transmitted to the offspring (figure 1.1). For any pair of loci, the alleles on the child chromosome come from two opposite parental chromosomes and are said to be recombinant when an odd number of crossovers occurred between them, and they come from the same parental chromosome when an even number of crossovers (including none) occurred between them. Figure 1.1 gives an example with zero and one crossover. Alleles at loci close to each other on a parental chromosome have a high probability of being transmitted together from the parent to the offspring, i.e. a low recombination probability. The recombination probability is an increasing function of the distance between the two loci. Distant loci and loci on different chromosomes have a recombination fraction of $\frac{1}{2}$. Loci with a probability of recombination $< \frac{1}{2}$ are said to be linked. Ott [44] provides a more detailed introduction to the concepts underlying linkage analysis. We elaborate on the notion of identity by descent and its central role in linkage analysis in section 2.2.

![Figure 1.1: Schematic illustration of a crossover between a pair of homologous chromosomes during the prophase of meiosis. At that point the chromosomes are duplicated to form chromatides linked at the centromere. Any pair of chromatides can participate in an exchange.](image)

The position of an unknown gene is inferred by detecting linkage between that gene and landmarks whose genome location is known called genetic markers. A genetic marker is a DNA polymorphism, meaning that it has two or more alleles with sufficiently high frequency in the population of interest. It is detected by a DNA probe binding a preferably unique DNA segment near the marker itself. The different alleles are detected by a biochemical assay. By extension of the notion of observed trait, the outcome of the assay
is called the marker phenotype. In the absence of error in the measurement procedure, the two observed alleles forming this phenotype coincide with the actual genotype of the marker.

Genetic markers are anchored to physical locations on the genome, and their order and the physical distances between them, in number of bases, can be determined from a physical map of the genome. For linkage analysis, a notion of distance related to the probability of recombination between loci in a meiosis is needed. The genetic distance $d$ is defined as the expected number of crossovers between two loci. Its unit is the Morgan (M) with the centiMorgan (cM) as a commonly used subdivision. Recombination fractions are estimated by the count of recombinant chromosome copies over the total number of chromosome copies produced by a sample of meioses. In humans the count of recombinants is inferred from the marker phenotypes of family members. Genetic distances cannot be estimated directly from empirical data, but only by fitting a model of the crossover process to the recombination counts. The function relating recombination fraction $\theta$ and genetic distance is called a map function $\theta = M(d)$. A Poisson model for the crossover process leads to the Haldane map function $\theta = \frac{1}{2}(1 - e^{-2d})$. An ordered set of markers and the genetic distances separating them form a genetic map. The goal of linkage analysis is to assign a position to unknown genes influencing a trait on a genetic map.

For a series of loci mapping to the same chromosome, each chromosome copy of an individual carries a series of alleles, referred to as a haplotype. The assay reading the phenotype of markers returns the alleles at individual loci without indication as to which allele belongs to which haplotype. Those genotypes are said to be unordered. The complete specification of genotype information includes the parental origin of the alleles in addition to their type, and is termed an ordered genotype. Multilocus ordered genotypes, or equivalently ordered haplotypes, cannot be observed in an isolated individual with standard biochemical methods, but can be derived from parental marker genotypes.

The fact that gene transmission is inferred from the phenotypes of parents and children means that families are needed to do linkage analysis. A set of relatives with known relationships forms a pedigree. Figure 2.1 gives an example of a pedigree drawing, with the males represented by squares and the females by circles. We adopt the convention of calling founders the individuals without parents in the pedigree and non-founders the
other pedigree members. The types of pedigrees collected for linkage analysis studies range from nuclear families with two children to the complete genealogies of isolated populations going back to the foundation of the population, for instance the populations of Saguenay-Lac-St-Jean in Québec and Iceland.

1.1.1 Notation

The variable $Y$ represents phenotypes throughout this thesis. $Y_D$ denotes a disease phenotype. All phenotypes in this thesis are categorical. A marker phenotype is associated to a particular locus and is indexed by the locus number $l$. $Y = (Y_{li}), l = 1, \ldots, L, i = 1, \ldots, I$ is the matrix of the marker phenotypes of all individuals in a pedigree at the marker loci considered in a particular analysis.

Unless stated otherwise genotypes are ordered and are represented by the letter $g$. The genotype of individual $i$ at locus $l$ is denoted $g_{li}$, and $g = (g_{li}), l = 1, \ldots, L, i = 1, \ldots, I$ is the matrix of multilocus marker genotypes of the pedigree members.

Genetic loci on the human genome are identified by a standard code consisting of the letter D followed by the chromosome number (or X or Y), the letter S and a locus-specific number. An example would be D1S123. The numbering does not reflect the physical order of the loci.

1.2 Outline

This thesis examines two general aspects of linkage analysis of genetic traits in large multi-generation pedigrees: the application of Markov chain Monte Carlo (MCMC) sampling methods to obtain approximations to linkage statistics using multiple genetic markers simultaneously with the aim to improve power compared to using a single marker, and the use of identity by descent information between pedigree members to detect disease genes. That methodology is then applied to the linkage analysis of glaucoma in a large pedigree from the Glaucoma Inheritance Study in Tasmania.

The thesis is organized as follows. In chapter 2 we define meiosis indicators, identity by descent and the relation between the two. We review the methods based on sharing
of genes IBD between affected pedigree members for detecting genetic linkage to a disease trait in pedigrees of arbitrary structure. A method for assessing significance of statistics measuring IBD is presented next. We then discuss the difficulties created by missing genetic marker information, introduce the notion of assigning a probability to IBD sharing patterns conditional on marker data and present the standard model of multilocus inheritance, a hidden Markov model with meiosis indicators as hidden variables. An explanation of the use of the conditional expectation of identity by descent sharing statistics given marker data under imperfect information concludes this chapter.

In chapter 3, we present a MCMC sampling approach to performing multilocus computations on large pedigrees. It combines the meiosis and locus Gibbs samplers developed by others and new types of Metropolis samplers. We also examine how marker data on the children in a nuclear family may create classes of states that do not communicate by applying only the meiosis sampler. Estimators of the Monte Carlo variance of the estimates based on the theory of time series and convergence diagnostics applied to the sequence of realizations of statistics computed from the states of the Markov chain are described next. This chapter also includes a study of the performances of diverse combinations of samplers in various proportions in term of bias and variance of estimates and computing time for test problems where the complexity of the computations is within the limits of what exact algorithms can handle.

The practical aspects and results of the linkage analysis of a large pedigree of the Glaucoma Inheritance Study in Tasmania are presented in chapter 4. We begin the chapter by reviewing the current knowledge on the genetics of glaucoma and presenting the pedigree analyzed and the available data. We then introduce IBD sharing statistics designed from the hypotheses on the genetics of the disease and the pedigree structure for linkage analysis of glaucoma in the pedigree, and evaluate the power of those statistics to detect glaucoma predisposing genes in the pedigree. The results of the MCMC computations of IBD sharing statistics using the data from 401 markers on the genome are then presented, starting with an assessment of the difficulty to reach convergence of the estimates on a problem of that size. Some regions of the genome are further investigated in the last sections of the chapter. Chapter 5 contains a summary of the thesis and concluding remarks.
The distinctive features of this thesis are:

- **Extension of the methods for MCMC sampling of meiosis indicators by introducing new samplers for meiosis indicators and combining multiple types of samplers in hybrid samplers.** The hybrid samplers can be readily applied to general pedigrees and marker data within some limits of size without the need to design data specific Markov chain steps that restricts the applicability of MCMC sampling on genotypes. Tuning of an hybrid sampler to improve sampling on a particular type of problem is simply done by varying the combination of samplers and the proportions of the included samplers. Guidelines for the choice of hybrid samplers are given based on an assessment of the performance of hybrid samplers in term of bias and variance of estimates in two test cases.

- **Estimation of the Monte Carlo variance of estimates and application of convergence diagnostics to the output of MCMC runs.** Variance estimation and convergence assessment have been generally ignored in the applications of MCMC to pedigree analysis.

- **Empirical investigation of the applicability of MCMC sampling on a large pedigree.** The usefulness of convergence diagnostics to detect failure of the MCMC sampler to sample the target distribution is investigated.

- **Design of IBD sharing statistics based on the pedigree structure and hypotheses on the genetic etiology of a disease.** The power of such statistics is compared to the power of a general purpose statistic.
Chapter 2

Genetic linkage analysis with pedigrees

Genetic linkage analysis locates genes with respect to genetic markers based on the principle that alleles at two nearby loci on the genome tend to be transmitted together from parent to offspring. We begin by presenting a coding scheme for gene transmission, the meiosis indicators. We then define the concept of identity by descent, show that meiosis indicators determine IBD sharing and explain how IBD sharing can be exploited to locate disease genes. A review of the allele sharing methods, either sharing by state or sharing by descent, proposed for general pedigrees follows. A simulation method for assessing significance of IBD sharing statistics is presented next.

In large pedigrees the inheritance information obtained from individual genetic markers is often incomplete. A hidden Markov model of multilocus inheritance is described to combine information from multiple linked marker loci. The modifications to IBD sharing statistics under incomplete information conclude this chapter.

2.1 The meiosis indicator representation of inheritance

The relevant information for linkage analysis is the grand-parental origin of the alleles that a parent passes on to his child through the process of meiosis. That information can be encoded in binary indicators.
For a given meiosis $i$ at locus $l$, define:

$$S_{li} = \begin{cases} 
0 & \text{if parent’s paternal allele is transmitted} \\
1 & \text{if parent’s maternal allele is transmitted}
\end{cases}$$

This coding scheme for the transmission of genes introduced to compute probabilities of gene descent by Donnelly [12] was applied to linkage analysis by Lander and Green [35] who termed the vector $S_l$ of all the indicators at one locus the inheritance vector. In this thesis we adopt the terminology of Thompson [59] and call $S_{li}$ a meiosis indicator. When the marker phenotypes allow us to unambiguously track the transmission of genes at a locus in a pedigree they determine the meiosis indicators, up to an inversion of the phase of the indicators for the meioses from a founder to his offspring. Figure 2.1 represents three loci where marker phenotypes of the members of a small pedigree determine the value of the meiosis indicators.

Figure 2.1: Marker phenotypes and meiosis indicators at three loci. In founders the phenotypes correspond to unordered genotypes, with alleles separated by commas. In non-founders the ordered genotypes inferred from the phenotypes are shown with alleles separated by a vertical bar. The meiosis indicators of the first child of each founder are arbitrarily set to 0.

Under the first law of Mendel the two parental alleles have equal probability of being transmitted to the child, hence
\[ P[S_{li} = 0] = P[S_{li} = 1] = \frac{1}{2} \]  

(2.1)

When a recombination occurs between two loci \( l \) and \( j \) the alleles transmitted to the child at the two loci come from opposite parental chromosomes and the indicator at locus \( l \) takes a value different from the indicator at locus \( j \). The probability of a recombination event is the recombination fraction \( \theta_{l,j} \):

\[ P[S_{li} = 0 | S_{ji} = 1] = P[S_{li} = 1 | S_{ji} = 0] = \theta_{l,j} \]

The meioses in the mother and the father of a child are independent. There is conclusive evidence that paternal and maternal genetic distances differ in humans and other species, however they will be assumed to be equal to the average of the two to simplify computations.

### 2.2 Identity by descent sharing and its role in linkage analysis

When a gene is replicated and passed on to an offspring, the new copy is said to be identical by descent (IBD) to the parent copy. Two genes that are copies of the same ancestral gene are also identical by descent. More generally, identity by descent extends to a set of genes that are copies of the same ancestral gene and the ancestor gene itself. Barring mutations, IBD genes are of same allelic type.

A pedigree structure represents the ancestry of individuals and therefore determines the possible paths of descent of genes. In linkage analysis the only familial relationships considered are those represented in some well circumscribed pedigree. Relationships outside that pedigree structure are ignored, so the genes of the pedigree founders are assumed distinct by descent.

Methods of linkage analysis based on IBD rely on the fact that affected relatives have a higher probability of sharing genes IBD at or near a locus influencing susceptibility to a disease than their probability to share at an unlinked locus under a wide range of models of the relationship between genotype and disease phenotype. Those models have in common an assumption of conditional independence between the phenotypes of different
individuals given genotypes and environmental factors influencing the disease. A strict form of that assumption rules out correlation between the environmental effects on relatives and conditions only on genotype.

**Assumption 1** *The disease phenotype of an individual is conditionally independent of the phenotypes and genotypes of his relatives given his multilocus genotype at all loci affecting his predisposition to the disease.*

\[ P[Y_D|g] = \prod_{i=1}^{I} P[Y_{Di}|g_i] \]

### 2.2.1 Illustration of the increased probability of IBD sharing between affected relatives

For a simplified illustration of this general principle, consider two unilineally related individuals, i.e. individuals connected to their common ancestor(s) via a single line of descent, affected by the same disease. Let \( IBD \) represent the event that the two individuals share an allele IBD and \( \phi_i, i = 1, 2 \) the event that individual \( i \) is affected by the disease. We will show that \( P[IBD|\phi_1, \phi_2] > P[IBD] \) when the susceptibility to the disease is influenced by a gene with two alleles \( A \) and \( a \) and the disease risk ratio \( r \) between individuals with genotype \( Aa \) or \( AA \) and individuals with genotype \( aa \) is strictly greater than 1.

**Proof.** The following notation and conditions will be used:

- The probability of \( IBD \) conditioning only on the familial relationship between the pair of relatives and Mendel laws is denoted \( \pi > 0 \).
- The baseline disease risk for genotype \( aa \) is \( f_0 > 0 \).
- The two individuals are from a population where allele \( A \) has frequency \( p > 0 \).

For simplicity we assume Hardy-Weinberg equilibrium at the disease gene locus, the assumption 4 of section 2.5 which implies marginal independence of the allelic types of any set of genes. We first establish that \( P[IBD|\phi_1, \phi_2] - \pi \) is proportional to \( P[\phi_1, \phi_2|IBD] - P[\phi_1, \phi_2|\overline{IBD}] \) and then show that \( P[\phi_1, \phi_2|IBD] - P[\phi_1, \phi_2|\overline{IBD}] > 0 \) when \( r > 1 \).
So first, we have:

\[
P[\text{IBD}|\phi_1, \phi_2] = \pi = \frac{\pi}{\pi_{[\phi_1, \phi_2]}} (P[\phi_1, \phi_2|\text{IBD}] - P[\phi_1, \phi_2])
\]

\[
= \frac{\pi}{\pi_{[\phi_1, \phi_2]}} (P[\phi_1, \phi_2|\text{IBD}] - [P[\phi_1, \phi_2|\text{IBD}] \pi + P[\phi_1, \phi_2|\neg \text{IBD}](1 - \pi)])
\]

\[
= \frac{\pi(1-\pi)}{\pi_{[\phi_1, \phi_2]}} (P[\phi_1, \phi_2|\text{IBD}] - P[\phi_1, \phi_2|\neg \text{IBD}])
\]

The expression for \( P[\phi_1, \phi_2|\text{IBD}] \) in terms of \( p \), \( r \) and \( f_0 \) is derived by conditioning on the event that the pair of individuals shares allele \( A \) or allele \( a \):

\[
P[\phi_1, \phi_2|\text{IBD}] = P[\phi_1, \phi_2|\text{share A, IBD}]P[\text{share A}|\text{IBD}]
\]

\[
+ P[\phi_1, \phi_2|\text{share a, IBD}]P[\text{share a}|\text{IBD}]
\]

\( P[\text{share A}|\text{IBD}] \) is the probability that the common ancestral gene is of type \( A \), i.e. \( p \). The conditional probabilities of being affected are:

\[
P[\phi_1, \phi_2|\text{share A, IBD}] = P[\phi_1|\text{share A, IBD}]P[\phi_2|\text{share A, IBD}] \text{ under assumption 1}
\]

\[
= rf_0 \cdot rf_0 = r^2 f_0^2
\]

\[
P[\phi_1, \phi_2|\text{share a, IBD}] = P[\phi_1|\text{share a, IBD}]P[\phi_2|\text{share a, IBD}] \text{ under assumption 1}
\]

\[
P[\phi_i|\text{share a, IBD}] = P[\phi_i|g_i = aa]P[g_i = aa|\text{share a, IBD}]
\]

\[
+ P[\phi_i|g_i = Aa]P[g_i = Aa|\text{share a, IBD}]
\]

\[
= f_0(1 - p) + rf_0 p
\]

Putting the pieces together, we get:

\[
P[\phi_1, \phi_2|\text{IBD}] = P[\phi_1, \phi_2|\text{share A, IBD}]P[\text{share A}|\text{IBD}]
\]

\[
+ P[\phi_1, \phi_2|\text{share a, IBD}]P[\text{share a}|\text{IBD}]
\]

\[
= r^2 f_0^2 p + (f_0(1 - p) + rf_0 p)^2(1 - p)
\]

\[
= [r^2 p + (1 - p + rp)^2(1 - p)] f_0^2
\]

If the two affected individuals do not share a gene IBD, their phenotypes and genotypes are independent by assumptions 1 and 4. The probability that either individual is affected is then given by

\[
P[\phi_1, \phi_2|\neg \text{IBD}] = P[\phi_1|\neg \text{IBD}]P[\phi_2|\neg \text{IBD}]
\]

\[
P[\phi_i|\neg \text{IBD}] = P[\phi_i|g_i \neq aa]P[g_i \neq aa|\neg \text{IBD}] + P[\phi_i|g_i = aa]P[g_i = aa|\neg \text{IBD}]
\]

\[
= rf_0(1 - (1 - p)^2) + f_0(1 - p)^2
\]
The joint probability is then

\[
P[\phi_1, \phi_2 | IBD] = P[\phi_1 | IBD] P[\phi_2 | IBD]
\]

\[= [r f_0 (1 - (1 - p)^2) + f_0 (1 - p)^2]^2
\]

\[= [r p (2 - p) + (1 - p)^2]^2 f_0^2
\]

To show that \( P[\phi_1, \phi_2 | IBD] - P[\phi_1, \phi_2 | \overline{IBD}] > 0 \), we expand the two probabilities in \( r \) (after dividing by \( f_0^2 \)). The terms of the polynomials are then proportional to the probability of \( \phi_1, \phi_2 \) given that 0, 1 or 2 of the individuals carry at least one allele \( A \), for the case where the individuals share an allele IBD and for the case where they do not. The demonstration is completed by subtracting the terms of same degree.

\[
(P[\phi_1, \phi_2 | IBD] - P[\phi_1, \phi_2 | \overline{IBD}]) / f_0^2 = r^2 [p + p^2 (1 - p)] + 2rp (1 - p)^2 + (1 - p)^3
\]

\[-\left\{ r^2 p^2 (2 - p)^2 + 2rp (2 - p) (1 - p)^2 + (1 - p)^4 \right\}
\]

\[= r^2 p (1 - p)^3 - 2rp (1 - p)^3 + p (1 - p)^3
\]

\[= (r - 1)^2 p (1 - p)^3 > 0 \text{ when } r > 1.
\]

\[
\square
\]

### 2.2.2 Identity by descent configuration

At a given locus the \( 2n \) genes of a set of \( n \) relatives fall into classes of genes identical by descent. Here we use the term gene in the sense of a piece of genetic information present in an individual. An IBD configuration specifies to which class of identical genes the two unordered genes of each individual belong. Leaving the genes unordered implies that paternal and maternal genes are indistinguishable. This is a common assumption in genetic models. The phenomenon of parental imprinting where the genetic effect of an allele depends on whether it was transmitted from the mother or from the father is the exception to that rule.

For a given ordering of the \( n \) individuals, Thompson [56] defined a unique labeling of the genes in an IBD configuration. The genes are ordered such that the maternal gene precedes the paternal gene of an individual. The maternal gene of the first individual receives the label 1. The other genes are then consecutively examined. If a gene is identical by descent with a gene preceding it, it receives the label of that gene. Otherwise, it receives
label \( j + 1 \), where \( j \) is the number of different labels assigned to the genes examined so far. Once all genes are labeled, pairs where the label of the maternal gene is greater than the label of the paternal genes are switched.

Consider the following example of two sibs and their aunt (figure 2.2). When the trio is ordered \((21, 22, 13)\), the seven possible IBD configurations between them given their familial relationships are:

\[
\begin{array}{ccc}
21 & 22 & 13 \\
1 & 2 & 1 \\
1 & 2 & 1 \\
1 & 2 & 1 \\
1 & 2 & 1 \\
1 & 2 & 1 \\
1 & 2 & 1 \\
1 & 2 & 1 \\
\end{array}
\]

The labeling would differ for another ordering of the trio, but the sharing relations described by each IBD configuration would remain the same.

Figure 2.2: Example of IBD configuration between two sibs and an aunt.

The collection of IBD configurations for a given set of \( n \) individuals that are compatible with their relationships in a given pedigree must be distinguished from the set of potential IBD configurations between \( n \) individuals. In the special case of 2 individuals, there
are 9 potential IBD configurations also known as condensed identity states (Jacquard [27]) and denoted $\Delta_1$ to $\Delta_9$. They enter in the computation of genetic correlations used in variance component analysis of quantitative traits measured on members of pedigrees.

**Proposition 1** *The IBD configuration at a locus is determined by meiosis indicators*

*Proof.* Assign distinct labels from 1 to $2f$ to the genes in the $f$ pedigree founders at some locus. Then let the genes be transmitted from parents to offsprings in successive generations, according to the vector of meiosis indicators at the locus. Next, collect the gene labels of the $n$ individuals whose genes are included in the IBD configuration. Genes with the same label are IBD and distinct by descent from genes with a different label, so the collection of gene labels specifies the IBD configuration. A unique labeling is then obtained by applying the procedure of Thompson [56] described above.

IBD sharing statistics are univariate measurements of aspects of IBD sharing that are expected to be markedly different at or near a disease locus compared to a random locus. They can be used to test the null hypothesis that no disease gene is linked to the locus where they are computed. This approach is advocated for genetic linkage detection because it does not require specification of a genetic model relating the disease phenotype to the genotype at an hypothesized disease locus. As a result, some IBD sharing statistics are found to be robust with respect to disease model. (Dudoit and Speed [13], Davis and Weeks [9]). Various methods proposed to exploit allele sharing between affected individuals to locate genes are reviewed in the next section, from measuring identity by state to deriving IBD sharing statistics from a genetic model for a trait.

**2.2.3 Review of allele sharing methods of linkage analysis in general pedigrees**

Allele sharing methods were initially designed for linkage analysis in nuclear families (Day and Simons [10]). Several extensions to arbitrary pedigree structures of varying size have been proposed in the past fifteen years.

The first of these extensions, the affected pedigree members (APM) method, is based on identity by state of alleles in affected individuals (Weeks and Lange [62]). The
statistic computed is the sum of the number of pairs of alleles identical by state. The mean and variance of the statistics under the null hypothesis of no linkage to a disease gene are computed from the probability of identity by descent between genes in pairs of affected individuals (the condensed identity states) and the allele frequencies in the population. The significance of the observed statistic is then assessed using a normal approximation or by simulation of the statistic null distribution.

Davis et al. [51] found the APM method to have low power due to the inclusion of alleles that are of the same state but nevertheless distinct by descent in the count of identical pairs and proposed counting pairs of allele identical by descent instead of merely identical by state in the SimIBD statistic. The simulated conditional distribution of the statistic given the marker phenotypes of unaffected individuals is used as a null distribution to reduce the dependence of the distribution on allele frequencies.

Both the APM and SimIBD methods can be applied to arbitrarily large pedigrees. They utilize a single marker at a time (an extension of the APM method adds the statistic at multiple markers but is not using the information from multiple markers jointly) and information extracted from a single marker is limited due to the factors explained in section 2.4.

Recognizing that the IBD configuration of the affected individuals contains all the information on their gene sharing Whittemore and Halpern [63] proposed to use it as basic IBD encoding. Different aspects of IBD sharing are quantified by defining score functions on IBD configurations. Whittemore and Halpern [63] defined two such functions:

\( S_{\text{pairs}} \) : Count of the number of pairs of alleles IBD between affected individuals in a pedigree.

\( S_{\text{all}} \) : Form \( 2^k \) vectors \( \omega \) by sampling one gene from each of the \( k \) affected individuals in a pedigree. For each vector compute the number of permutations \( h(\omega) \) of the genes preserving the identity by descent relations between the genes in the vector. If there are \( u \) distinct by descent alleles with multiplicity \( b_1, \ldots, b_u \) summing to \( k \) in vector \( \omega \), then \( h(\omega) = \prod_{j=1}^{u} b_j! \) and \( S_{\text{all}} = \frac{1}{k} \sum_{\omega} h(\omega) \).
The statistic $S_{pairs}$ generalizes to arbitrary family relationship between affecteds the count of all pairs of alleles IBD between affected siblings, a statistic that has been studied and applied for several years (Suarez and Van Eerdewegh [54]). With affected siblings, it has the property of being the score statistic in the recombination parameter $\theta$ between a chromosomal location and a disease locus to test the null hypothesis $\theta = \frac{1}{2}$, independently of the genetic model (Dudoit and Speed [13]). Kruglyak et al. [32] present results of simulation studies on small pedigrees indicating that $S_{all}$ is more powerful than $S_{pairs}$ under a range of genetic models.

Other IBD scoring functions were introduced by Sobel and Lange [52] who apply them on large pedigrees. McPeek [39] also proposed different scoring functions, justifying them from their equivalence to likelihood ratio statistics under limiting genetic models.

The joint probability of the disease phenotypes and an IBD configuration $B$ at a disease susceptibility locus is determined by a genetic model specifying penetrances of genotypes and genotype frequencies and by the pedigree structure and the laws of inheritance. When the IBD configuration and disease phenotypes are the observed data, the likelihood ratio to test the null hypothesis $H_0$ that the locus where the IBD configuration is observed is unlinked to the disease locus against the alternative $H_1$ that it coincides with the disease locus is:

\[
LR(Y_D, B) = \frac{P_{H_1}[Y_D, B]}{P_{H_0}[Y_D, B]} = \frac{P_{H_1}[Y_D | B]P[B]}{P[Y_D]P[B]} = \frac{P_{H_1}[Y_D | B]}{P[Y_D]}
\]

McPeek [39] derives Taylor expansions of $P_{H_0}[Y_D|B]$ around a boundary point of the parameter space of a two allele genetic model, for instance around disease allele frequency equal to 0 or penetrance ratio between high and low risk alleles equal to 1. The coefficients of the first order term in the expansion are functions of patterns of identity by descent between affected individuals. When the parameter in function of which $P_{H_0}[Y_D|B]$ is expanded tends to its limit, higher order terms become negligible and the function of IBD in the first order term is equivalent to the likelihood ratio. Under that limiting model the test based on the corresponding IBD sharing function is locally most powerful.
2.3 Assessing significance of identity-by-descent sharing test statistics

When the IBD configuration of the affected individuals in a pedigree is observed, the statistical significance of any IBD sharing statistic $Z$ is measured against its distribution conditional on disease phenotype under the null hypothesis that a locus is unlinked to any disease gene. Under that null hypothesis, the meiosis indicators at one locus are mutually independent and independent of the disease phenotype. All meiosis vectors are therefore equiprobable and by computing the value of $Z$ for each one the null distribution of $Z$ can be computed. When the number of meiosis vectors is too large, enumeration can be replaced by random sampling of a sufficient number of realizations of $S_l$.

2.3.1 Genomewide significance

The analysis of a genome scan involves testing the null hypothesis of no linkage at a large number of loci. The critical values of the test has to be adjusted so that the probability of rejecting the null hypothesis at any point on the genome when no locus is associated with the disease does not exceed the prespecified significance level $\alpha$.

In the following development the autosomal genome is treated as a continuous interval of length $G$. The approximation could be refined by considering instead that the genome is made of 22 segments of different lengths (the chromosomes) but the effect on the critical values would be minimal. Under that model a test statistic follows a stochastic process in continuous distance along the genome. For large samples of pedigrees with the same structure, for instance nuclear families with an affected sib pair, the process can be approximated by an Ornstein-Uhlenbeck process and critical values derived from excursion probabilities (Feingold [16], Lander and Kruglyak [34]).

With a single large pedigree of arbitrary structure no simple process could approximates the process of IBD sharing statistics satisfactorily. Simulation of the process is the preferred solution. The approach outlined here has been described by Durham and Feingold [14].
Every meiosis is represented as a binary process defined over the length of the genome taking the value 0 where the segment transmitted to the child is of paternal origin and 1 where the segment is of maternal origin. The process switches value at crossover points. Crossovers are assumed to occur following a Poisson process with rate 1 in each meiosis. The meiosis processes are initialized at one end of the genome by sampling the value of the meiosis indicators independently from distribution 2.1. The meiosis processes are generated independently over the length of the genome.

We consider two functions of the IBD sharing statistic process:

- The maximum of the IBD sharing statistic over all the values of the meiosis vector seen over the length of the genome.

\[ Z_{\text{max}}^{(c)} = \max_{0 \leq t \leq G} \{Z_t\} \]

- The maximum over the values of \( Z \) observed at \( d \) cM intervals along the genome, to mimic a genome scan with that spacing between markers.

\[ Z_{\text{max}}^{(d)} = \max_{t \in \{d, \ldots, G\}} \{Z_t\} \]

\[ Z_{\text{max}}^{(c)} \leq Z_{\text{max}}^{(d)} \] and for small values of \( d \) the two statistics will be close. The critical value \( c_{\alpha} \) for \( Z_{\text{max}}^{(c)} \) say, is defined as \( \inf \{z : P[Z_{\text{max}}^{(c)} \geq z] \leq \alpha \} \).

The null distributions of \( Z_{\text{max}}^{(c)} \) and \( Z_{\text{max}}^{(d)} \) are approximated by repeating the simulation of the crossover process in the whole pedigree a large number of times.

### 2.4 Missing inheritance information in pedigrees

In practice, the available marker phenotypes do not determine the meiosis indicators in a pedigree as they do in figure 2.1. The incompleteness of the information extracted from marker phenotypes has three main sources:

**Unavailable individuals** The marker phenotypes of deceased ancestors and family members that could not be reached or refused to participate to the study are missing.
Uninformative meioses The grand-parental origin of a transmitted allele cannot be determined. This occurs when a parent has an homozygous phenotype, making his paternal and maternal alleles indistinguishable, or when two parents and their child have the same marker phenotype so that it is not possible to know which allele the child got from which parent.

Marker assay failures The biochemical assay used to observe marker phenotypes fails to produce an output some fraction of the time, or the phenotypes read are inconsistent with the rules of Mendelian inheritance and must be discarded.

Even with incomplete marker phenotypes the IBD configuration between specific subsets of individuals may sometimes be determined unambiguously. In general however multiple IBD configurations are consistent with the incomplete marker phenotypes. In large pedigrees spanning several generations the unavailability of all the ancestors in the top generations results in a wide range of possible meiosis indicator realizations and IBD configurations. In the large pedigree from a study of glaucoma in Tasmania analyzed in chapter 4.2.1, no marker phenotypes were observed on 119 individuals out of 246, including all individuals in the top three generations of the six generation pedigree. In individuals on which marker phenotypes were measured, the mean proportion of phenotypes that could not be determined from the measurement procedure was 19%. In addition, an average of 23% of individuals were homozygous at the observed marker loci, creating uninformative meioses when the individuals have offspring in the pedigree.

2.4.1 Inferring identity by descent

When marker information is incomplete, probabilistic models based on the laws of inheritance can be used to assign probabilities to meiosis indicators outcomes or IBD sharing patterns. Different approaches have been used. Davis et al. [51] infer the IBD status of every pair of alleles in affected individuals from all available marker phenotypes in the pedigree at a single marker locus using a recursive algorithm. When IBD status is ambiguous, the conditional IBD probability the pair of alleles \((a_i, a_j)\), \(P[a_i \text{ IBD to } a_j | Y]\), is computed. When not all pedigree members have observed marker phenotypes, only an approximation of the values of \(P[a_i \text{ IBD to } a_j | Y]\) is computed.
The more general approach adopted in this thesis is to compute probabilities of IBD configuration at a genetic map location conditional on the marker phenotypes at one or several markers. The intuitive advantages of using multiple markers are that the information missing at one marker may be filled in by the data at nearby markers, and observing alleles of the same type at a string of markers in related individuals connected by unavailable ancestors is stronger evidence of IBD than observing alleles of the same type at a single marker. There is empirical evidence that the power of linkage analysis is increased by using multiple markers.

Curtis and Sham [8] compute the IBD configuration between pairs of individuals using the pedigree peeling algorithm for recursive probability computations over pedigrees (see section 3.4.1) as implemented in the LINKAGE computer package [37]. The method is straightforwardly extended to include multiple linked markers for the computation of IBD probabilities at one of the loci, but is subject to the restrictive limits on the number of loci that can be handled with the pedigree peeling algorithm.

Computation of the probability of any IBD configuration conditional on marker phenotypes at multiple loci is made possible if the problem is reformulated as the computation of conditional distributions of meiosis indicators $S$ given multilocus marker phenotypes, and the value of $S$ mapped to IBD configurations (section 2.2.2). In the following section we describe the model under which inference on meiosis indicators are done. It was first proposed by Lander and Green [35].

2.5 A hidden Markov model of multilocus inheritance

The marker phenotypes of the individuals in a pedigree and the variables encoding the inheritance of their genes at several loci form a large system of random variables with a complex structure when all potential dependencies are taken into consideration. The rules of Mendelian genetics specify local dependencies between the genes of parents and children in the pedigree that simplify computations in large pedigrees at a single locus. In order to make computations involving many loci possible, simplifying assumptions on the relationships between loci are required. They are here formulated in the meiosis indicators representation.
Assumption 2  Conditional independence between the phenotype at a marker locus and
phenotypes and meiosis indicators at other marker loci given the meiosis indicators at the
current locus.

\[ P[Y_i|S, \{Y_k, k \neq l\}] = P[Y_i|S_i] \]

The hidden implication of this assumption is the independence of the marker geno-
types between marker loci in the founders of the pedigree. This can be seen by expanding
the above expression to the founder genotypes \(fg\).

\[
P[Y_i|S, \{Y_k, k \neq l\}] = \sum_{fg} P[Y_i, fg|S, \{Y_k, k \neq l\}]
= \sum_{fg} P[Y_i|S, fg, \{Y_k, k \neq l\}] P[fg|S, \{(fg_k, Y_k), k \neq l\}]
\times P[\{fg_k, k \neq l\}|S, \{Y_k, k \neq l\}]
= \sum_{fg} P[Y_i|S, fg] P[fg|\{fg_k, k \neq l\}] P[\{fg_k, k \neq l\}|\{(S_k, Y_k), k \neq l\}]
\]

The simplification of the first two terms on the line above results from the fact that \((S, fg)\)
fully specifies gene inheritance at all loci and therefore no additional information on \(Y_i\) or
\(fg_i\) is provided by the marker phenotypes at other loci. The marginal independence between
\(S\) and \(fg\) then makes conditioning on \(S\) in the second term unnecessary. With the founder
genotypes explicitly represented, assumption 2 is reformulated as \(P[Y_i|S, fg] = P[Y_i|S_i, fg_i]\)
and \(P[fg_i|\{fg_k, k \neq l\}] = P[fg_i]\), showing the independence between the genotypes at dif-
ferent loci in the founders sampled from the population. This strong assumption is referred
to as linkage equilibrium in genetic terminology. Loci distant enough (maybe \(\geq 10\) cM) will
tend to be in linkage equilibrium in a large population with no stratification present. The
marginal formulation of the assumption follows from summing over the founder genotypes:

\[
P[Y_i|S, \{Y_k, k \neq l\}] = \sum_{fg_i} P[Y_i|S_i, fg_i] P[fg_i] \sum_{\{fg_k, k \neq l\}} P[\{fg_k, k \neq l\}|\{(S_k, Y_k), k \neq l\}]
= P[Y_i|S_i]
\]
**Assumption 3** First order Markov dependence between meiosis indicators at successive marker loci.

\[ P[S_l|\{S_k, k \neq l\}] = P[S_l|S_{l-1}, S_{l+1}] \]

This assumption implies that crossover events occur following a Poisson process on the chromosome and that crossovers in non-overlapping intervals are independent. In actual meioses, the probability of a crossover tends to be lower than predicted from the Poisson model in an interval around a crossover point, a phenomenon known as positive interference. The Poisson or no-crossover interference model is adopted for the computational simplification it affords. Distances derived from this model are close to the actual ones when the distances are not too small (≥ 5 cM say). Assumptions 2 and 3 define a hidden Markov model (Baum and Petrie [1]) with Y the observed data and S the hidden states.

When there are unobserved founder genotypes \( f_g \) as happens in multigeneration pedigrees, the computation of \( P[Y_l|S_l] \) involves summing over their values weighted by their probability \( P[f_g] \). An estimate of \( P[f_g] \) is then needed. Typically the available marker data from unrelated individuals in the population allow us only to estimate allele frequencies, and the efficient algorithm to perform the sum over \( f_g \) described in appendix A requires that the distribution \( P[f_g] \) be expressed as a product of allele frequencies. For those reasons, it is convenient to make the further assumption of Hardy-Weinberg equilibrium at the marker loci, although it is not required for the hidden Markov model.

**Assumption 4** Hardy-Weinberg equilibrium: At a given locus the probability of the genotypes of a set of individuals in a population is the product of the frequencies of the alleles within each individual and across the individuals.

\[ P[g_l] = \prod_i P[g_{l,im}]P[g_{l,ip}] \]

where \( g_{l,im} \) and \( g_{l,ip} \) represent the paternal and maternal alleles respectively.

This assumption implies that the alleles forming the genotypes are sampled independently from the population gene pool. When the matings in the population are independent of the alleles carried by the mates at a locus, the population is large and there is no selection, mutation or immigration then the genotype frequencies reach Hardy-Weinberg
equilibrium in one generation and remain stable in successive generations. Those conditions are however not met in real human populations.

The genotype probabilities computed under the Hardy-Weinberg and linkage equilibrium assumptions are an approximation to the actual probabilities that may be good in an homogeneous population for markers with enough distance between them and with alleles that have no phenotypic effect that could influence mate choice. The substitution of the allele frequencies entering in the computation of $P[Y_l|S_l]$ and the recombination fractions entering in the computation of $P[S_l|S_{l-1}, S_{l+1}]$ by estimates, usually obtained from external data is another level of approximation. The IBD configuration probabilities derived under the model are functions of the marker data $Y$ that quantify the evidence contained in the data in favor of an IBD configuration.

The vectors of phenotypes $Y_l$ and meiosis indicators $S_l$ at each locus can be depicted as vertices of a graph where edges represent dependencies between variables. The graph shown in figure 2.3 has the general form of hidden Markov model graph.

![Figure 2.3: Graphical representation of the hidden Markov model formed by the marker phenotype (filled circles) and meiosis indicator vectors (empty circles) at a sequence of linked loci.](image)

2.5.1 Inferring conditional distribution of meiosis indicators

From the joint probability of meiosis indicators $S$ and observed marker data $Y$ under the above hidden Markov model, the conditional distribution of meiosis indicators at
a locus $l$ is derived as:

$$P[S_l | Y] = \frac{\sum_{S_{i,k\neq l}} P[S,Y]}{\sum_{S} P[S,Y]}$$

The marker data at all loci are therefore used in inferring the meiosis indicators distribution.

When the dimension of $S_l$ is small (up to $\approx 20$ meioses after dimension reductions made possible by symmetries in the pedigree), $P[S_l | Y]$ can be computed exactly for all values of $S_l$ with a complexity linear in the number of loci using a version of the forward-backward algorithm for HMM of Baum and Petrie [1] known as the Lander-Green algorithm (Lander and Green [35]). Speed was improved and the meiosis vector space reduced by Kruglyak et al. [32] and Gudbjartsson et al. [23] but the algorithm still suffers from an exponential growth in complexity with the number of meioses. In larger pedigrees, approximation methods are required. In the next chapter, we develop a Markov chain Monte Carlo approach to sample observations from $P[S_l | Y]$ that are then used to approximate linkage statistics.

### 2.6 IBD sharing statistics under incomplete information

IBD sharing statistics at a locus are not observable when the IBD configuration between the affected individuals of a pedigree is not uniquely determined by the marker phenotypes. Whittemore and Halpern [63] proposed to estimate the unobservable value of a statistic $Z$ by its conditional expectation given multilocus marker phenotypes $Y$:

$$\bar{Z}_l = E_l[Z | Y] = \sum_s Z_l(s) P[S_l = s | Y]$$

It would be logical to assess the significance of $\bar{Z}$ against its null distribution, but this presents difficulties. First, $\bar{Z}$ is a function of the multilocus marker phenotypes $Y$. The joint distribution of $Y$ depends on the joint distribution of multilocus genotypes of the founders $fg$. In a general model for that distribution, the population frequency of every combination of multilocus genotypes of a couple (mating type) would be a free parameter. Population samples are generally too small to estimate all the parameters of that high dimensional model. When only allele frequencies can be estimated with some accuracy, the joint distribution of $fg$ is expressed as a product of those frequencies under the independence assumptions of linkage equilibrium between marker loci (as in assumption 2) and Hardy-Weinberg equilibrium at each marker locus (assumption 4). $\bar{Z}$ reflects information on IBD
even when the assumptions are not exactly true, but relying on the same assumptions to compute its distribution may lead to incorrect inferences, as the product distribution may be a poor approximation to the true distribution of $fg$, especially if the markers are tightly linked.

Obtaining the distribution of $Z$ is also computationally burdensome because of the difficulty of evaluating $Z$. On small pedigrees one execution of the Lander-Green algorithm already takes on the order of several seconds. On large pedigrees the computation of $Z$ using MCMC methods like those described in chapter 3 is a daunting task requiring hours of computing time. It is then not feasible to evaluate $Z$ for the hundreds of realizations of $Y$ needed to obtain a good approximation to the distribution of $Z$.

The third difficulty with the distribution of $Z$ is that it depends on genomic location. Besides adding to the computational burden by requiring to compute a different distribution for every locus considered, the location heterogeneity makes impossible to study the behavior of the statistic on a genome-wide scale. This prevents the derivation of critical values for the maximum of $Z$ over the whole genome.

For those reasons, Kruglyak et al. [32] proposed to use the distribution of $Z$ under complete information to assess significance of $Z$. The variance of $Z$ is larger than the variance of $Z$ as can be seen from the following expression for conditional variance:

$$Var[E[Z|Y]] = Var[Z] - E[Var[Z|Y]] \leq Var[Z]$$

For large numbers of pedigrees, where the distribution of the sample means of $Z$ and $Z$ tend to normal distributions, this result guarantees that p-values for $Z$ derived from the distribution of $Z$ are conservative. With just one or a few pedigrees, the distribution of $Z$ is not Gaussian and its tail could be larger than the tail of the distribution of $Z$. Empirical results reported in section 4.8.1 indicate that this is unlikely and in fact the approximation can be very conservative. The distribution of $Z$ is generally used in this thesis as null distribution for $Z$.

A problem is that the distribution of $Z$ has a much more limited number of support points than the distribution of $Z$. Many IBD statistics take positive integer values, while
$Z$ can take fractional values. We propose spreading the probability point mass in the tail of the distribution of $Z$, $F_Z$, over real intervals. It assumes that the tail of $F_Z$ decreases exponentially. Between the support points $x_{j-1}$ and $x_j$, the survival of the continuous distribution $S_C = 1 - F_C$ is defined by the exponential interpolation of the survival $S_Z = 1 - F_Z$:

$$S_C(u) = \exp \left\{ \log S_Z(x_j) - (\log S_Z(x_{j-1}) - \log S_Z(x_j)) \frac{x_j - u}{x_j - x_{j-1}} \right\}, \quad x_{j-1} \leq u \leq x_j$$

$S(u)$ is the one-sided p-value of $u$. Interestingly, when $Z$ is estimated by the sample mean of a sample of realizations of $Z$, the p-value obtained from the above method is equal to the geometric mean of the p-values of each realization of $Z$, a p-value computation approach proposed by Camp et al. [4].

Kong and Cox [31] approach the problem of incompleteness of the information by fitting a one-parameter alternative model relating the chosen IBD sharing statistic $Z$ to the conditional distribution $P[S_l|Y]$. They propose two forms of alternative models:

**Linear model**

$$P_\delta[S_l = s] = P[S_l = s] \{1 + \delta(Z(s) - \mu)/\sigma\}, \quad 0 \leq \delta \leq \frac{\sigma}{\mu - \inf Z}$$

**Exponential model**

$$P_\delta[S_l = s] = \frac{\exp \{\delta(Z(s) - \mu)/\sigma\}}{\sum_{s'} \exp \{\delta(Z(s') - \mu)/\sigma\}}, \quad 0 \leq \delta$$

where $\mu = E[Z]$ and $\sigma = \sqrt{\text{Var}[Z]}$. With multiple pedigrees, the parameter $\delta$ is common to all pedigrees, and a weight $\gamma$ multiplying $(Z(s) - \mu)/\sigma$ can be assigned to each pedigree. The log-likelihood of the observed data $(Y_D, Y)$, $l(\delta)$ is then written as a function of $P_\delta[S_l = s]$ and maximized with respect to $\delta$. The null hypothesis $\delta = 0$ is tested using the log-likelihood ratio statistic $\lambda = 2(l(\hat{\delta}) - l(0))$. Relying on the likelihood theory result that the distribution of $\lambda$ converges to a $\chi^2$ distribution with one degree of freedom as the number of pedigrees becomes large, Kong and Cox [31] suggest the use of that distribution to assess significance of $\lambda$. Assuming the asymptotic distribution is a good approximation, the level of the test is accurate regardless of the level of information from the markers. With very few pedigrees, the same authors find that the $\chi^2(1)$ tends to have a shorter upper tail than the distribution of $\lambda$, the discrepancy being worse under the linear model than the exponential model.
Unfortunately, only the linear model is applicable to large pedigrees due to its property that the likelihood depends on \((Y_D, Y)\) only through \(E_l[Z|Y]\), which can be approximated without computing the entire distribution \(P[S_l|Y]\). The requirement to compute the entire distribution \(P[S_l|Y]\) to fit the exponential model make it intractable for large pedigrees. The method of Kong and Cox [31] is therefore unsuitable for a single large pedigree.
Chapter 3

Markov chain Monte Carlo sampling of meiosis indicators

The exponential growth in complexity of the computation of $P[S|Y]$ with the number of meioses limits the size of pedigrees handled by the Lander-Green algorithm and its enhanced variants. In pedigrees too large for exact computation, samples of realizations of $S$ drawn from $P[S|Y]$ can be used to compute approximations to linkage statistics. We explain how the estimates are computed and why MCMC is a practical method for sampling from $P[S|Y]$. The remaining of the chapter is devoted to the description of the various types of samplers applied to construct Markov chains on $S$, the presentation of empirical tools for convergence monitoring and the evaluation of the precision and accuracy of various sampler combinations on test problems. The software implementation of the MCMC samplers is described in appendix B.

The statistics used in linkage analysis can be formulated as expectations of functions of $S$ with respect to $P[S|Y]$. This is true both for likelihood ratio statistics under a parametric model for the trait and for IBD sharing statistics (Kruglyak et al. [32]). The IBD sharing statistic case is treated in section 2.6. Provided that observations can be sampled from $P[S|Y]$, the conditional expectation

$$
\tilde{Z}_l = E_l[Z|Y] = \sum_s Z_l(s)P[S_l = s|Y]
$$

is estimated by $\hat{Z} = \frac{1}{N} \sum_r Z(S^{(r)})$. 
Direct sampling of independent realizations from $P[S|Y]$ would be the preferred approach if it were feasible. However, sampling from a high dimensional discrete distribution is practical only under the condition of being able to evaluate the probability of any possible outcome in the sample space. Given the way the genetic model is defined, $P[S|Y]$ can only be expressed indirectly in terms of transmission probabilities $P[S]$ and phenotype probabilities conditional on transmission $P[Y|S]$ using Bayes theorem:

$$P[S|Y] = \frac{P[Y,S]}{P[Y]} = \frac{P[Y|S]P[S]}{\sum_{S'} P[Y|S']P[S']}$$

The denominator is the probability of the observed marker phenotypes, which is computed by summing over all values of $S$, a summation whose complexity is the reason for turning to sampling methods in the first place.

The situation where one defines a function $\pi^*$ over a sample space $\mathcal{X}$ and is interested in the probability distribution $\pi$ defined over the same space and proportional to $\pi^*$ arises in many scientific areas, and for multidimensional spaces the normalizing constant $C = \sum \pi^*$ such that $\pi = \pi^*/C$ is uncomputable. Metropolis et al. [40] introduced the idea of defining a Markov chain on the state space $\mathcal{X}$ with $\pi$ as its stationary distribution and transition probabilities that depend only on ratios $\pi_i/\pi_j = \pi_i^*/\pi_j^*$. This means that only the unnormalized function $\pi^*$ is required to generate the Markov chain. After the chain has been run long enough to reach its stationary distribution, realizations of the chain are samples from $\pi$.

The Markov chain must satisfy some conditions for the method to work. All the states in $\mathcal{X}$ must communicate, forming a single irreducible class. This is obviously needed to sample the whole state space with a single chain. Irreducibility must be established for every problem for which a Markov chain is constructed, and this is often a non trivial task. The states of the chain must also be recurrent, meaning that the chain returns to all states with probability 1, and aperiodic, which can be described by saying that the chain has no cyclic behavior and may visit any state at any time point. The recurrence condition will be satisfied on finite state spaces as long as there is no absorbing state from which the chain can’t get out. Collectively, the above conditions define the property of ergodicity. Ergodicity implies that the chain has a unique stationary distribution. The ergodic theorem insures
that estimates of the expectation of functions defined over $X$ converges to the expectation, for instance that $\tilde{Z}$ defined above converges to $E[Z|Y]$. A more rigorous description of Markov chain theory concepts required for MCMC is given by Roberts [48].

Metropolis et al. [40] defined the transition matrix $P$ of the Markov chain indirectly via a proposal transition matrix $Q$. In the original Metropolis algorithm $Q$ is symmetric; Hastings [24] generalized the proposal to an arbitrary stochastic matrix. The Metropolis-Hastings algorithm can be summarized as follows:

- Start from any initial state $x_0$.
- Given $x_{t-1} = i$, set $x^* = j$ with probability $q_{ij}$, i.e. from the $i^{th}$ row of $Q$
- Accept $x^*$, i.e. put $x_t = j$ with probability $\min(1, H_{ij})$, $H_{ij} = \frac{q_{ji}}{q_{ij}}$ otherwise put $x_t = i$.

The $H_{ij}$ are known as the Hastings ratios. The transition probabilities $p_{ij}$ of this two-stage chain are:

$$p_{ij} = \begin{cases} q_{ij}H_{ij} & i \neq j \\ q_{ii} & i = j \end{cases}$$

and satisfy the detailed balance condition

$$\pi_ip_{ij} = \pi_jp_{ji} \forall i, j$$

This in turn implies that $\pi$ is the stationary distribution of the chain since

$$\sum_j \pi_jp_{ji} = \sum_j \pi_i \pi_{ij} = \pi_i.$$

The Gibbs sampler (Geman and Geman [20]) is a special case of the Metropolis-Hastings algorithm that applies to multivariate probability distributions. We present it in the context of its application to the sampling of meiosis indicators. At each step, a subset of the variables is sampled conditional on the value of all other variables. With the meiosis indicator matrix $S$, two types of subsets are the rows $S_{*,i}$ corresponding to the indicators of a meiosis at all loci (meiosis sampler, section 3.1) and the columns $S_l$ corresponding to all meioses at a locus (locus sampler, section 3.4).
Using the conditional distribution as proposal gives acceptance probabilities that are all equal to 1. For sampling of all the indicators at one locus (whole locus sampler of section 3.4), the Hastings ratio is:

\[ H(S, S^*) = \frac{P[S^*_1, S^*_2, Y]/q_l(S^*_1, S^*_2)}{P[S_1, S_2, Y]/q_l(S_1, S_2)} = \frac{P[S^*_1, S^*_2, Y]/P[S^*_1|S_2, Y]}{P[S_1, S_2, Y]/P[S_1|S_2, Y]} = 1 \]

MCMC methods have been applied for some time to approximate probabilities and expectations in pedigree analysis. Most of the early developments involved sampling of ordered genotypes (see Lin [38] for a review). Thompson [58] first proposed to use meiosis indicators as the latent variables. This latter framework, that we adopted, presents multiple advantages. First, the space of meiosis indicators is much smaller than the space of ordered genotypes. Second, meiosis indicators do not always constrain the allelic types of genes since the allelic types of the transmitted genes are not specified, so that sometimes a Markov chain defined on meiosis indicators is irreducible where a Markov chain defined on ordered genotypes is not. Third, the identity by descent configurations between relatives are easier to extract from meiosis indicators than ordered genotypes.

### 3.1 Whole meiosis Gibbs sampler

The Lander-Green algorithm, a version of the forward-backward algorithm, computes probabilities on the HMM formed by marker phenotype data and meiosis vectors at multiple linked loci described in section 2.5 with a complexity linear in the number of loci. Thompson and Heath [60] took advantage of this property to devise an efficient scheme to sample the indicators of a single meiosis at all loci. The indicators of all meioses except the one being sampled are fixed, allowing a forward-backward algorithm to be applied on a vector of binary variables.

The update of one meiosis proceeds in two steps. In the forward recursion of the algorithm the probability distribution of the indicator at each locus is computed conditional on other meiosis indicators and marker phenotypes from the first up to the current locus. In the backward recursion the meiosis indicators are iteratively sampled in reverse order.
Forward recursion  The forward recursion is defined as follows:

\[
\alpha_1(S_{1,i}) = P[Y_1|S_1]P[S_1,i] \propto P[Y_1, S_1]
\]

\[
\alpha_l(S_{l,i}) = \alpha_l(S_{l-1,i})P[S_{l,i}|S_{l-1,i}]P[Y_l|S_l] \propto P[Y_1, \ldots, Y_l, S_l|\{S_{1,k}, \ldots, S_{l-1,k}, k \neq i\}]
\]

The term \(P[S_{l,i}|S_{l-1,i}]\) is the 2 \(\times\) 2 matrix

\[
\begin{bmatrix}
1 - \theta_{l-1} & \theta_{l-1} \\
\theta_{l-1} & 1 - \theta_{l-1}
\end{bmatrix}
\]

An efficient algorithm to compute \(P[Y_l|S_l]\) introduced by Sobel and Lange [52] and Kruglyak et al. [32] and considerations relative to its implementation are presented in appendix A.

Backward recursion  The last term of the forward recursion \(\alpha_L(S_{L,i})\) is normalized to get the distribution of \(S_{L,i}\) given phenotype data and meiosis indicators at other meioses \(P[S_{L,i}|Y, S_{L*, k \neq i}]\) and \(S_{L*, i}^*\) is sampled from that distribution. Other indicators in \(S_{*,i}\) are sampled iteratively in backward order. Suppose \(S_{L+1,i}^*, \ldots, S_{L,i}^*\) have been sampled. Then \(S_{l,i}^*\) is sampled from

\[
P[S_{l,i}|S_{l+1,i}^*, \ldots, S_{L,i}^*, \{S_{l,k}, k \neq i\}, Y] = \frac{\alpha_l(S_{l,i})P[S_{l,i}|S_{l+1,i}^*]}{\alpha_l(0)P[S_{l,i} = 0|S_{l+1,i}^*] + \alpha_l(1)P[S_{l,i} = 1|S_{l+1,i}^*]}
\]

The end result is a new realization of \(S_{*,i}\).

3.2 Non-communicating classes of states in nuclear families with the meiosis Gibbs sampler

Observed genotypes impose constraints on the genes of relatives that can and cannot be identical by descent. These constraints can create classes of meiosis indicator states that do not communicate when meioses are updated one at a time. In this section we study the constraints created by the genotypes of the children of a nuclear family. We only consider the case where the genotypes of the parents are unobserved, as it is often the unknown parental origin of the alleles that creates sets of non-communicating states. As pointed out by Thompson and Heath [60], when a sampler is irreducible at each individual
locus the whole sampler is irreducible. This results from the fact that the combination of valid meiosis indicator states at different loci produces valid multilocus states provided the recombination fractions between the loci are strictly positive. We can therefore restrict our attention to a single locus when establishing irreducibility.

We first derive the number of possible joint genotypes of a sibship such that the genotypes of the sibs are all distinct. The same constraints on meiosis indicators applying to sibs with identical genotypes, siblings with duplicate genotypes can be removed to reduce a sibship to one where all sibs have distinct genotypes while preserving the same classes of communicating meiosis indicator states. We then characterize every set of states with sub-classes that do not communicate by single meiosis flips.

The genotypes of sibs are independent draws from the $m$ possible genotypes formed by sampling one gene from each of the two parents. Unordered parental genotypes are denoted $g_1g_2 \times g_3g_4$ where $g_i$ takes an allelic type denoted by a capital letter. There are $m = 4$ possibilities when the parents carry 4 distinct alleles $AB \times CD$ or the two parents possess an allele of the same type which is distinct from the other alleles $AB \times AC$. We have $m = 3$ when the two parents have identical heterozygous genotypes $AB \times AB$ and $m = 2$ when one parent is homozygous and the other heterozygous $AA \times BC$ or $AA \times AB$. The trivial case $m = 1$ occurs when the two parents are homozygous $AA \times BB$ or $AA \times AA$ and provides no information on the meiosis indicators.

Forming a joint genotype of a sibship of size $k$ with no repeated individual genotype is the same as choosing $k$ items without replacement out of $m$. The number of possible joint genotypes is therefore given by

$$N_g = \binom{m}{k}$$

Table 3.1 enumerates all joint genotypes with no duplicate for $m = 4$ and $k = 2, 3, 4$. The joint genotypes for $m = 3$ are obtained by substituting $C$ by $B$ in the column with three allelic types and the joint genotypes for $m = 2$ are obtained by substituting $C$ by $A$ in either column and collapsing identical joint genotypes. Different sets of meiosis indicator states are compatible with different joint genotypes. Some of the joint genotypes generate one of three state sets containing sub-classes that do not communicate by flipping only one meiosis
indicator at a time and a reference to the appropriate state set is included in Table 3.1. A description of each state set and an illustration of a joint sibship genotype generating it follows.

<table>
<thead>
<tr>
<th>$k = 4, \ N_g = 1$</th>
<th>4 allelic types</th>
<th>State set</th>
<th>3 allelic types</th>
<th>State set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 3, \ N_g = 4$</td>
<td>$AB; AD; BC; CD$</td>
<td>3</td>
<td>$AA; AB; AC; BC$</td>
<td>3</td>
</tr>
<tr>
<td>$AB; AD; BC$</td>
<td>1</td>
<td>$AA; AB; AC$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$AB; AD; CD$</td>
<td>1</td>
<td>$AA; AB; BC$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$AB; BC; CD$</td>
<td>1</td>
<td>$AA; AC; BC$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$AD; BC; CD$</td>
<td>1</td>
<td>$AB; AC; BC$</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Joint genotypes of a sibship such that the genotypes of the sibs are all distinct and there are 4 possible individual genotypes. The numbers refer to the state sets described in the text. “Other” refers to a state set that is not described because it is irreducible with single meiosis flips.

1. Two sibs with two different alleles in a sibship of size two or three (Figure 3.1).

![Figure 3.1: Two sibs with different alleles.](image)

Observing two sibs with no allele in common implies that the two sibs share no gene IBD, and have opposite meiosis indicators, as listed in Table 3.2. To move between
states it is necessary to flip the paternal or the maternal indicators of both sibs simultaneously.

<table>
<thead>
<tr>
<th>11 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 11</td>
</tr>
<tr>
<td>01 01</td>
</tr>
<tr>
<td>10 10</td>
</tr>
<tr>
<td>11 00</td>
</tr>
</tbody>
</table>

Table 3.2: Meiosis vectors of the two sibs on figure 3.1 consistent with their observed genotypes.

2. Three sibs with genotypes such that each child has one allele in common with each of his two siblings (figure 3.2).

![Figure 3.2: Three sibs with genotypes such that each child has one allele in common with each of his two siblings.](image)

This case was studied by Thompson and Heath [60] who list the compatible states of the meiosis indicators in their table 3. Only two of the pairs of identical-type alleles are IBD. The third pair is distinct by descent, each parent transmitting one copy. The valid unordered parental genotypes are $AB \times AC$, $AB \times BC$ and $AC \times BC$. Flipping both meiosis indicators of each individual simultaneously is the step involving the minimal number of meioses that makes all valid states communicate.
3. Four sibs with genotypes such that each child has one allele in common with two of his siblings and none with his third sibling (figure 3.3).

![Figure 3.3: Four sibs with genotypes such that each child has one allele in common with two of his siblings and none with his third sibling.](image)

This results in the two parents having genotypes that are not found in the children. In the particular example depicted in figure 3.3, the unordered parental genotypes are \( AC \times BD \). The compatible states of the meiosis indicators are given in table 3.3. Flipping the paternal or the maternal indicators of all sibs simultaneously steps between valid configurations, but divides the state space into two non-communicating classes. A combination of steps where all the sibs with concordant indicators have their indicators flipped and all the sibs with discordant indicators have their indicators flipped are needed to define an irreducible chain, or a combination of one of those steps with the flip of paternal or maternal indicators.

This section covered only the instances of non-communicating classes with the meiosis Gibbs sampler within a nuclear family. Updating multiple meioses simultaneously produces irreducible samplers in those cases. One scheme is covered in the next section. With larger multigeneration pedigrees the constraints on the meiosis indicators become much more complex and steps updating indicators within a nuclear family do not guarantee an irreducible sampler. In those cases the locus sampler of section 3.4 is needed.
Table 3.3: Meiosis vectors of the four sibs on figure 3.3 consistent with their observed genotypes.

3.3 Updating all meioses between a parent and his children

In some of the cases presented in section 3.2, updating all the meioses between a parent and his children results in an irreducible chain in a nuclear family. The computational complexity of sampling from the distribution over the $2^k$ different values of the vector of $k$ indicators between a parent and his $k$ children grows exponentially with the number of children, and many of the $2^k$ values have probability 0 due to the constraints imposed by the marker phenotypes, especially with large number of children. It is more efficient to sample from a limited set of configurations of the indicators. In this section we introduce the parental sampler which extends to multiple loci the Metropolis step that Sobel and Lange [52] call rule $T_1$. This simple proposal gives an irreducible chain in the three cases presented in section 3.2. It flips all the indicators of the meioses from the parent $p$ to his kids at a subset of consecutive marker loci $L_0, \ldots, L_k$. The proposal is formally:

$$S^*_{li} = \begin{cases} 
1 - S_{li} & \text{if } i \in \{j : p \in \text{parents}(j)\}, \ l \in \{L_0, \ldots, L_k\} \\
S_{li} & \text{else.}
\end{cases}$$

To minimize the number of new recombinations introduced by the flip, only subsets of consecutive markers including the first or the last marker in the map are considered, that is $\{1, \ldots, l\}$ and $\{l, \ldots, L\}$, $l = 1, \ldots, L$. 

\[
\begin{array}{cccc}
11 & 12 & 13 & 14 \\
01 & 00 & 11 & 10 \\
01 & 11 & 00 & 10 \\
00 & 10 & 01 & 11 \\
00 & 01 & 10 & 11 \\
10 & 00 & 11 & 01 \\
10 & 11 & 00 & 01 \\
11 & 10 & 01 & 00 \\
11 & 01 & 10 & 00 \\
\end{array}
\]
For a subset extending from marker 1 to \( l \), the ratio of the state probabilities is the ratio of the joint meiosis indicator-marker genotype probability at each of the updated loci times the difference between the Hamming distance \( h(\cdot, \cdot) \) between loci \( l \) and \( l + 1 \) in the current state \( S \) and in the proposed state \( S^* \).

\[
\frac{P[S^*, Y]}{P[S, Y]} = \frac{P[S^*_1, Y_1] \cdots P[S^*_l, Y_l]}{P[S_1, Y_1] \cdots P[S_l, Y_l]} \theta^{h(S^*_l, S^*_{l+1}) - h(S_l, S_{l+1})} (1 - \theta)^{M - (h(S^*_l, S^*_{l+1}) - h(S_l, S_{l+1}))}
\]

Note that \( S^*_{l+1} = S_{l+1} \). The proposal is symmetric, so \( S^* \) is accepted with probability

\[
\min \left( \frac{P[S^*, m]}{P[S, m]}, 1 \right)
\]

to preserve detailed balance.

The proposal is applied in a predetermined or random order to the meioses from every parent in the pedigree so that all the meioses in the pedigree are involved in exactly one step in each cycle.

Thomas et al. [55] present an elegant formulation of the update of the \( k \) indicators of the children of an individual involving an auxiliary binary variable \( X_l \) at each of the \( L \) loci. For a given parent, the value 0 of this variable is associated to \( S_l \) and 1 to \( S_l^* \). The variables \( X_1 \) to \( X_L \) are then sampled from their conditional distribution given \( Y \) and \( \{S_{*,i}, i \notin \{j : p \in \text{parents}(j)\}\} \) in the same way as meiosis indicators in the meiosis Gibbs sampler. The setup as a Gibbs sampler has the advantage of eliminating the acceptance/rejection step. Thomas et al. [55] also suggest sampling jointly the auxiliary variables for the two parents in a nuclear family.

### 3.4 Whole locus Gibbs sampler

As there exist a recursive algorithm, the Lander-Green algorithm, linear in the number of loci \( L \) to compute the probability of the phenotype data observed on a pedigree, a similar algorithm exploits the tree structure of pedigrees without loops to compute the probability with a complexity linear in the number of individuals in the pedigree. This algorithm suffers from an exponential growth in \( L \), but it nonetheless suggests an efficient meiosis indicator sampling scheme: updating \( S_l \), the meiosis indicators at locus \( l \), at once,
just as $S_{*i}$ is updated at once by the meiosis sampler applied to meiosis $i$.

An important advantage of the whole locus Gibbs sampler is that it is irreducible, provided that all recombination fractions between loci are strictly positive. This is because all meiosis indicators at a locus are sampled jointly, insuring that all compatible single locus states can be reached, and the positive probability of recombination between loci allows any pattern of recombination events between the meiosis indicators at different loci. This theoretical guarantee of irreducibility does not prevent that in practice some transitions introducing multiple recombinations between linked loci have very low probability, possibly causing the chain to remain in a sub-region of the space for a very long time.

We begin by reviewing the principles of that recursive probability computation algorithm on trees called *pedigree peeling*. The whole locus Gibbs sampler is then constructed by describing the three steps it involves: peeling of the pedigree at one locus conditional on the flanking marker loci, reverse peeling and generation of the new realization of $S_l$.

### 3.4.1 Pedigree peeling

Recursive computation of probabilities over the individuals in a pedigree without loops is made possible by the conditional independence structure of ordered genotypes. For any individual $i$ in such a pedigree, define the set of individuals “above” him, denoted $\{k: k < i\}$, as the set of pedigree members related to him via his parents, (including ancestors and siblings and their descendents, spouses and spouses’ ancestors) and the set of individuals “below” him, noted $\{k: k > i\}$, as the set of pedigree members related to him via his children, (including spouses and spouses’ ancestors, descendents, descendents’ spouses and their ancestors). Then the ordered genotypes of the individuals above $i$ are conditionally independent from the ordered genotypes of the individuals below $i$ given $i$’s genotype. That is:

$$g_{\{k:k<i\}} \perp g_{\{k:k>i\}} | g_i$$

Exploiting that property, one needs to find an ordering of individuals to minimize the dimension of the joint genotype distribution at any step of the probability computation. In a pedigree without loop, orderings of the individuals exist that are optimal in the sense
that the dimension of the probability distributions never exceeds 3, the dimension of a joint
genotype of a mother-father-child trio of individuals. Such an ordering can be obtained as
follows: pick a peripheral nuclear family, i.e. one connected to the rest of the pedigree by a
single individual called a pivot. Children of the nuclear family other than the pivot (if the
pivot is a child) are assigned the next positions in arbitrary order. If the pivot is a parent,
his spouse receives the next number, otherwise the parents take the next two positions
after the children in arbitrary order. Then the family is removed from the pedigree. This
process is repeated until there remains a single nuclear family, whose members receive the
last positions in the ordering, the children first and the parents last.

The computation of \( P(Y) \) by peeling incorporates the phenotype data of consecutive
nuclear families into \( P[Y_1, \ldots, Y_i, g_i] \), for \( i \) belonging to the subsequence of pivots in the
ordering by summing over the probability terms involving genotypes in the pivot’s nuclear
family whose members come just before him, an operation referred to as “clipping” the
nuclear family. The pivot is then treated like any member of the next family, except that
its genotype probability is replaced by \( P[Y_1, \ldots, Y_i, g_i] \). Elston and Stewart [15] developed
the algorithm to clip nuclear families onto a parent, and Lange and Elston [36] extended
it to clip nuclear families onto a child pivot. More details on pedigree peeling are given in
chapter 4 of the book by Thompson[57].

The idea of applying the pedigree peeling algorithm to perform locus by locus
Gibbs sampling of genotypes and meiosis indicators conditional on the meiosis indicators
at flanking loci was first stated by Kong [30], who did not elaborate on the details of the
computations involved. The whole locus Gibbs sampler was subsequently implemented by
Heath [25] in the program Loki to perform Bayesian analyses of quantitative trait loci.

Since no conditional independence structure between the meiosis indicators at a
locus can be exploited for efficient computation, a whole locus Gibbs sampler on a space of
meiosis indicators requires first to generate a realization of ordered genotypes. This involves
two steps: pedigree peeling at one locus conditionally on the meiosis indicators at flanking
loci and sampling ordered genotypes in reverse order of peeling. The ordered genotypes are
then converted into a realization of the meiosis indicators at the locus.
3.4.2 Pedigree peeling at one locus conditional on meiosis indicators at flanking loci.

For the purpose of the locus sampler, the intermediate terms of the peeling of the pedigree at a locus \( l \) involving the joint genotypes of mother-father-child trios \( g_{lc}, g_{lm} \) and \( g_{lp} \) are saved and the summation is halted before summing over the genotype of the last individual in the peeling order. Furthermore, pedigree peeling has to be adapted to compute the probability of the meiosis indicators at loci \( l - 1 \) and \( l + 1 \) jointly with the probability of the marker data at locus \( l \).

The only terms of the pedigree peeling computation that are affected by the inclusion of the meiosis indicators at flanking loci are the transmission probabilities from father and mother to offspring. To explain the required modifications, we refer to figure 3.4 depicting the transmission of a segment of chromosome to a child \( c \) from his father \( p(c) \) and his mother \( m(c) \). The subscript \( cp \) refers to the paternal chromosome of the child and \( cm \) to the maternal, so that \( S_{l,cp} \) is the indicator in the paternal meiosis and \( g_{l,cp} \) is the type of the allele in the child genotype that comes from his father at locus \( l \).

The conditional probability of the genotype of the child \( g_{lc} \) given his father and mother genotypes at locus \( l \) only, \( P[g_{lc}|g_{l,p(c)}, g_{l,m(c)}] \), is the product of the probabilities of the paternal transmission, \( P[g_{l,cp}|g_{l,p(c)}] \), and the maternal transmission, \( P[g_{l,cm}|g_{l,m(c)}] \), since the maternal and paternal transmissions are independent under the first law of Mendel. Considering the paternal transmission from an heterozygous father, i.e. a father with alleles of distinct types \( g_{l,p(c)p} \neq g_{l,p(c)m} \), the event \( g_{l,cp} = g_{l,p(c)p} \) implies that the father paternal allele is transmitted to the child, i.e. \( S_{l,cp} = 0 \) and conversely for the maternal allele. When the father is homozygous, i.e. his alleles are of the same type \( g_{l,p(c)p} = g_{l,p(c)m} \) then the type of the allele transmitted to the child provides no information on the grand-parental origin of the allele \( S_{l,cp} \). The same holds for the maternal transmission. The joint distribution of allele types and meiosis indicators is summarized in table 3.4.

The probability that the meiosis indicator at a flanking locus takes the value 0 or 1 conditional on the event that the paternal or maternal allele is transmitted at locus \( l \) is a function of the recombination probability \( \theta \) between the two loci and is independent
Table 3.4: Joint probabilities of meiosis indicators and allele identity events at one locus.

<table>
<thead>
<tr>
<th>Events</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_l, cp = g_l, p(c) \neq g_l, p(c) m$</td>
<td>1 0</td>
</tr>
<tr>
<td>$g_l, cp = g_l, p(c) m \neq g_l, p(c) p$</td>
<td>0 1</td>
</tr>
<tr>
<td>$g_l, cp = g_l, p(c) p = g_l, p(c) m$</td>
<td>$\frac{1}{2} \frac{1}{2}$</td>
</tr>
</tbody>
</table>

Figure 3.4: Diagram of gene transmission from parent to offspring.

The last four terms are similar expressions; the first one is given below:

$$
P[g_l, S_{l-1}, cp, S_{l-1}, cm, S_{l+1}, cp, S_{l+1}, cm | g_l, p(c), g_l, m(c)]
\begin{align*}
&= P[g_l, cp | g_l, p(c)] P[S_{l-1}, cp | g_l, cp, g_l, p(c)] P[S_{l+1}, cp | g_l, cp, g_l, p(c)] \\
&\quad \times P[g_l, cm | g_l, m(c)] P[S_{l-1}, cm | g_l, cm, g_l, m(c)] P[S_{l+1}, cm | g_l, cm, g_l, m(c)]
\end{align*}
$$

The last four terms are similar expressions; the first one is given below:

$$
P[S_{l-1}, cp | g_l, cp, g_l, p(c)] = 
\begin{cases} 
\theta_{l-1}^{S_{l-1}, cp} (1 - \theta_{l-1})^{1 - S_{l-1}, cp} & \text{if } g_l, cp = g_l, p(c) \neq g_l, p(c) m \\
(1 - \theta_{l-1})^{S_{l-1}, cp} \theta_{l-1}^{S_{l-1}, cp} & \text{if } g_l, cp = g_l, p(c) m \neq g_l, p(c) p \\
\frac{1}{2} & \text{if } g_l, cp = g_l, p(c) p = g_l, p(c) m
\end{cases}
$$
3.4.3 Sampling of the ordered genotypes

When reaching the last individual in the peeling order, \( I \), the algorithm returns the joint probability of his possible genotypes at locus \( l \), the meiosis indicators at flanking loci and the marker data at locus \( l \), \( P[g_{l,I}, S_{l-1}, S_{l+1}, Y_l] \). The genotype of that last individual can then be sampled from its distribution conditional on marker data at \( l \) and meiosis indicators at \( l - 1 \) and \( l + 1 \):

\[
P[g_l | S_{l-1}, S_{l+1}, Y_l] = \frac{P[g_{l,I}, S_{l-1}, S_{l+1}, Y_l]}{\sum_{g_{l,I}} P[g_{l,I}, S_{l-1}, S_{l+1}, Y_l]}
\]

Sampling of the genotypes of each individual then proceeds in reverse peeling order, conditioning on the genotypes already sampled. This gives rise to two possible cases when it comes to processing a nuclear family: either the pivot individual (whose genotype was sampled first) is a parent or else it is a child. When the pivot is a parent, the genotype of its spouse is sampled first. Without loss of generality, assume the pivot is the father. Let \( an(i) \) denote the set comprising individual \( i \) and his ancestors and \( des(i) \) denote the set of his descendents. The genotype of the mother is sampled from:

\[
P[g_{l,m} | g_{l,p}, S_{l-1}, S_{l+1}, Y_{l,1}, \ldots, Y_{l,m}] \propto P[Y_{des(m)} S_{l-1,des(m)}, S_{l+1,des(m)} | g_{l,m}, g_{l,p}] P[g_{l,m}, S_{l-1, an(m)}, S_{l+1, an(m)}, Y_{l, an(m)}]
\]

The two terms were saved during the peeling operation. If \( m \) is a founder, then the second term simplifies to \( P[g_{l,m}, Y_{l,m}] \). Next, the genotype of each of the couple’s children is sampled from:

\[
P[g_{l,c} | g_{l,m}, g_{l,p}, S_{l-1}, S_{l+1}, Y_{l,1}, \ldots, Y_{l,m}] \propto P[g_{l,c}, S_{l-1,c}, S_{l+1,c} | g_{l,m}, g_{l,p}] P[Y_{l,c} | g_{l,c}] P[S_{l-1,des(c)}, S_{l+1,des(c)}, Y_{l,des(c)} | g_{l,c}]
\]

The third term was saved during the peeling of the pedigree. If the child has no descendent in the pedigree, it has value 1.

When the pivot is a child in the nuclear family, the genotype of one of his parents, say his father, is sampled first from:

\[
P[g_{l,p} | g_{l,c}, S_{l-1}, S_{l+1}, Y_{l,1}, \ldots, Y_{l,p}] \propto \sum_{g_{l,m}} P[g_{l,p}, S_{l-1, an(p)}, S_{l+1, an(p)}, Y_{l, an(p)}] P[g_{l,m}, S_{l-1, an(m)}, S_{l+1, an(m)}, Y_{l, an(m)}] P[Y_{l,des(p)} | g_{l,c}, S_{l-1,des(p)} | g_{l,m}, g_{l,p}] P[g_{l,c}, S_{l-1,c}, S_{l+1,c} | g_{l,m}, g_{l,p}]
\]
The first three terms come from the peeling computations. The first two simplify to $P[g_{l,p}, Y_{l,p}]$ (respectively $P[g_{l,m}, Y_{l,m}]$) when the father (respectively the mother) is a founder. The third term equals 1 when the parents have no other child than the pivot. The next step is to sample the mother’s genotype conditional on the father’s from the conditional distribution:

$$P[g_{l,m}|g_{l,c}, g_{l,p}, S_{l-1}, S_{l+1}, Y_1, \ldots, Y_m]$$

$$\propto P[Y_{l,des(m)}\setminus c, S_{l-1,des(m)}\setminus c, S_{l+1,des(m)}\setminus c|g_{l,m}, g_{l,p}]$$

$$\times P[g_{l,m}, S_{l-1,an(m)}, S_{l+1,an(m)}, Y_{l,an(m)}]P[g_{l,c}, S_{l-1,c}, S_{l+1,c}|g_{l,m}, g_{l,p}]$$

The first and second terms simplify in the special cases mentioned above.

### 3.4.4 Sampling of the meiosis indicators

Once the ordered genotypes $g_l$ of all the individuals in the pedigree have been sampled using reverse peeling, it remains to convert the ordered genotypes into meiosis indicators. For informative meioses, i.e. when the parent is heterozygous, the type of the transmitted allele determines whether the paternal or maternal allele is transmitted (see table 3.4). For uninformative meioses, i.e. when the parent is homozygous, the meiosis indicator $S_{l,i}$ is sampled conditionally on $S_{l-1,i}$ and $S_{l+1,i}$. The conditional probability that the paternal indicator of individual $c$, $S_{l,cp}$, is equal to 1 can be expressed as:

$$P[S_{l,cp} = 1|S_{l-1,cp}, S_{l+1,cp}] \propto \theta_{l-1}^{1-S_{l-1,cp}}(1 - \theta_{l-1})^{S_{l-1,cp}}\theta_{l}^{1-S_{l+1,cp}}(1 - \theta_{l})^{S_{l+1,cp}}$$

The resulting vector of meiosis indicators $S_l$ is a realization from $P[S_l|Y_l, S_{l-1}, S_{l+1}]$.

### 3.5 Generating an initial configuration of meiosis indicators

In order to start the Markov chain, one needs to generate a valid initial state. In the MCMC analysis of pedigrees, this can prove a difficult problem. The approach followed here is to use the whole locus Gibbs sampler to generate the initial state of the meiosis indicators $S$. This was first suggested by Heath [26] who sampled ordered genotypes. The same technique is extended to meiosis indicators by adding the extra step of generating $S$ from the realization of the ordered genotypes as in section 3.4.4.

The procedure can be outlined as follows:
1. Start at an arbitrary seed locus \( l \). Sample \( g_l \) from \( P[g_l|Y_l] \) using reverse peeling. Generate \( S_l \) from \( P[S_l|g_l] \).

2. For loci \( h = l+1, \ldots, L \) sample \( g_h \) from \( P[g_h|Y_h, S_{h-1}] \) using reverse peeling conditional on meiosis indicators at the preceding locus. Generate \( S_h \) from \( P[S_h|g_h] \).

3. For loci \( h = l-1, \ldots, 1 \) sample \( g_h \) from \( P[g_h|Y_h, S_{h+1}] \) using reverse peeling conditional on meiosis indicators at the following locus. Generate \( S_h \) from \( P[S_h|g_h] \).

The joint initial state \( S \) at all loci is therefore sampled from the probability distribution

\[
P[S_l|Y_l]P[S_{l-1}|Y_{l-1}, S_l] \cdots P[S_1|Y_1, S_2]P[S_{l+1}|Y_{l+1}, S_l] \cdots P[S_L|Y_L, S_{L-1}] \tag{3.1}
\]

This can be compared to sampling from the stationary distribution \( P[S|Y] \) following the same sequence of loci.

\[
P[S|Y] = P[S_l|Y_l]P[S_{l-1}|Y_1, \ldots, Y_{l-1}, S_l] \cdots P[S_1|Y_1, S_2] \\
P[S_{l+1}|Y_{l+1}, \ldots, Y_L, S_l] \cdots P[S_L|Y_L, S_{L-1}]
\]

Sampling from \( P[S|Y] \) is not practical since \( P[S_l|Y_l] \) is not computationally tractable. But one can see that (3.1) is an approximation to the full conditional distribution \( P[S|Y] \) where \( P[S_l|Y_l] \) is substituted by \( P[S_l|Y_l], P[S_{h-1}|Y_1, \ldots, Y_{h-1}, S_h] \) by \( P[S_{h-1}|Y_{h-1}, S_h] \), \( h = 1, \ldots, l-1 \) and \( P[S_{h+1}|Y_{h+1}, \ldots, Y_L, S_h] \) by \( P[S_{h+1}|Y_{h+1}, S_h] \), \( h = l+1, \ldots, L \).

### 3.6 Reinitializing the chain in a Metropolis step

Moving between distant high probability regions of the space of meiosis indicators by stepping through regions of low probability with the meiosis and locus samplers can be difficult to achieve when the number of meioses and the number of loci is large, resulting in slow mixing of the sampler as we will see in section 4.6.

One may consider running multiple chains from different initial configurations. Gelman and Rubin [19] proposed to first locate the modes of the distribution, and then to sample a set of initial values from a distribution with the same modes but a wider dispersion.
The multiple chains started from those initial values should together cover all major modes of the distribution. One criticism of this approach is that it is often not possible to find the modes of the distribution on which MCMC sampling is applied (Geyer [21], Cowles and Carlin [6]). This is the case with the conditional distribution of meiosis indicators \( P[S|Y] \).

Instead of starting independent chains, the reinitialization can be set up as the proposal of a Metropolis step. Provided that the proposed configurations are well spread over the space of valid \( S \), one can hope that all major modes of the distribution will be visited and the chain will spend a time proportional to their probability in each of them. We call this Metropolis step a chain restarting step.

Based on the initialization scheme described in section 3.5, we set up \( L \) Metropolis proposals, each one having a different locus as its seed locus. The proposal distribution \( q_l(.) \) for locus \( l \) is the sampling distribution (3.1) and is independent of the current state of \( S \). The probability \( q_l(S) \) can be re-expressed as

\[
q_l(S) = \frac{P[S_l|Y_l]P[S_{l-1}|Y_{l-1}, S_l] \ldots P[S_1|Y_1, S_2]P[S_{t+1}|Y_{t+1}, S_t] \ldots P[S_L|Y_L, S_{L-1}]}{P[Y_t]P[Y_{t-1}|S_t] \ldots P[Y_1|S_2]P[Y_{t+1}|S_t] \ldots P[Y_L|S_{L-1}]}
\]

From a current state \( S \), the Metropolis step is accomplished as follows:

1. Sample \( S^* \) from \( q_l(.) \).
2. Compute the Hastings ratio

\[
H(S, S^*) = \frac{P[S^*, Y]/q(S^*)}{P[S, Y]/q(S)}
\]

\[
= \frac{P[S^*, Y]}{P[S, Y]} \frac{P[Y_t]P[Y_{t-1}|S_t] \ldots P[Y_1|S_2]P[Y_{t+1}|S_t] \ldots P[Y_L|S_{L-1}]}{P[Y_t]P[Y_{t-1}|S_t] \ldots P[Y_1|S_2]P[Y_{t+1}|S_t] \ldots P[Y_L|S_{L-1}]}
\]

\[
= \frac{P[Y_{l-1}|S^*_l] \ldots P[Y_1|S^*_2]P[Y_{t+1}|S^*_t] \ldots P[Y_L|S^*_{L-1}]}{P[Y_{l-1}|S_l] \ldots P[Y_1|S_2]P[Y_{t+1}|S_t] \ldots P[Y_L|S_{L-1}]}
\]
3. Accept $S^*$ with probability $\min(H(S, S^*), 1)$.

The quantities $P[Y_h|S^*_{h+1}], h = 1, \ldots, l - 1$ and $P[Y_h|S^*_{h-1}], h = l + 1, \ldots, L$ are
by-products of the conditional pedigree peeling computation required to sample $S^*_h$ at loci
1, $\ldots, l-1, l+1, \ldots, L$. Computation of the corresponding probabilities for the current state
$S$ is also performed by peeling. The $L$ proposals $q_1(\cdot), \ldots, q_L(\cdot)$ can be applied in turn to
propose a variety of states. Since the meiosis vector of each locus is sampled all at once like
with the locus sampler, the chain restarting sampler shares its property of being irreducible.

However, two factors contribute to make $H$ low. The first is related to the proposal
distribution. Suppose chain restarting is the only type of step applied, that the current state
has been sampled from $q_h(\cdot)$ and that $S^*$ is being sampled from $q_j(\cdot), j > h$. Then, for
$h < l < j$, $P[S_l|Y_l, S_{l+1}]$ will tend to be much smaller than $P[S^*_l|Y_l, S^*_{l+1}]$ because $S_l$ was
sampled from $P[.|Y_l, S_{l-1}]$ conditional on the meiosis vector at the preceding locus, while
$S^*_l$ was sampled from $P[.|Y_l, S^*_{l+1}]$. When other types of steps are performed between the
chain restarts, the chain tends to move to a state with higher probability under $P[S|Y]$ than
the one sampled at the last reinitialization, but that may still have low probability under
$q_j(\cdot)$. A second factor reducing $H$ comes to play here. At that point the proposed state $S^*$
is likely to have lower probability than $S$, making the probability ratio $P[S^*, Y]/P[S, Y]
low. Experience shows that the chain restarting sampler has non-negligible acceptance
probability only on problems of moderate size. An empirical evaluation is presented in
section 3.8.

3.7 Monitoring convergence of estimates

In the application of the MCMC methodology, two notions of convergence of er-
godic Markov chains are of interest. The first type of convergence occurs when the Markov
chain approaches its stationary distribution, in term of variational distance, forgetting its
initial state. The transient phase between the start of the chain and the time it reaches
stationarity is referred to as the burn-in period. Realizations of the chain during that period
may introduce a bias in the estimates computed from the chain, hence the importance of
determining the time when the chain has reached stationarity to discard the observations
before that time.
When the expectation of functions defined over the state space is of interest, a second notion of convergence of the Markov chain is the convergence in probability of the empirical mean of the functions to their expectations, which is guaranteed by the ergodic theorem. The problem here is to determine the number of realizations to sample from the chain once it has reached its stationary distribution in order to obtain the desired precision.

Both types of convergence can be studied at different levels. A theoretical approach to convergence to stationarity consists in deriving bounds on the variational distance between the distribution of the chain after a finite time and its stationary distribution. Doing this with large state spaces requires some regular structure. The constraints on the space of meiosis indicators imposed by marker data are highly irregular and vary widely from one set of observations to the next, making the study of their theoretical convergence impractical and problem specific. Convergence of estimates to their expectation is related to convergence to stationarity and poses similar difficulties. Empirical assessment of convergence by analysis of the output of MCMC runs is more generally applicable, but lacks the rigor of theoretical results. We present a number of methods applied to the MCMC linkage analysis application described in this chapter.

Convergence to a high-dimensional distribution is difficult to measure and even more to visualize. Furthermore, the disk space required to store every realization of every variable may well exceed the available capacity. For those reasons, unidimensional summary statistics are most commonly monitored, with the hope that they capture the essential features of the distribution over the state space. When the latent variables are continuous, monitoring is often applied to the marginal distribution of each variable (Gelman [18]). This can’t be done with binary indicators. We opt to monitor functions defined over the state space whose expectations are the quantities to be estimated, another common choice.

For most of the datasets on which we applied MCMC sampling of meiosis indicators, a single run of the chain was performed. This was done to maximize the length of the runs that could be obtained with limited computing time. As pointed out by Geyer [21], a single long run of a chain yields more realizations from the stationary distribution for the same computing time than multiple chains after removing the burn-in period from each.
Running multiple chains started from different initial states is useful to detect convergence problems provided that the initial values are sampled from a distribution with higher dispersion than the target distribution (Gelman and Rubin [19]). In our case, we do not have the knowledge of the target distribution $P[S|Y]$ required to construct a distribution over-dispersed compared to it. Running supplementary chains started from arbitrarily chosen initial states may still provide a confirmation of failure to converge to the stationary distribution if the estimates from different runs are significantly different. Absence of discrepancy between runs does not however guarantee convergence because the initial states may be close to each other in terms of steps of the Markov chain.

The notion underlying both the convergence to the stationary distribution and the convergence of estimates to their expected value is the mixing speed of the chain, or the number of steps before a future realization of the chain is independent of the current one. Assessment of mixing speed is more fundamental than deciding on a burn-in time. A chain started from stationarity could give estimates with a large bias after a finite number of steps if it does not sample the whole state space extensively. The method of section 3.5 which conditions the starting values on subsets of the data should give starting values near a mode of the distribution when the dimension of the problem is within the range where fast mixing can be achieved with the samplers described in this chapter. In those cases the initial transient phase will be short and a short burn-in period will be sufficient. Cases where a long burn-in would be needed are also cases where the chain is too short to give accurate estimates, and this will be revealed by examining the sample. For all MCMC runs reported in this thesis, a preset burn-in phase representing between 1% and 5% of the run was discarded.

### 3.7.1 Convergence diagnostic tools

A multitude of methods have been proposed to diagnose both the convergence of the chain to its stationary distribution and the convergence of estimates to their expectation. Cowles and Carlin [6] and Brooks and Roberts [2] review a number of them. Most cannot be applied to assess convergence of the runs of our MCMC samplers for different reasons. Many require sampling multiple starting states from a distribution more dispersed than the target
distribution. Others estimate the density of the distribution of the sampled variables and do not apply to binary random variables. Some require the chain to have specific features that our samplers do not possess. This leaves few tools applicable to the present application.

A useful graphical tool to evaluate the mixing speed of a Markov chain is a cusum plot of a univariate statistic over the length of a run (Yu and Mykland [65]). The observed cusum, or partial sum, is defined as

\[ U_t = \sum_{r=1}^{t} [Z(S^{(r)}) - \hat{\mu}] \]

where \( \hat{\mu} = \sum_{r=1}^{N} Z(S^{(r)})/N \). Long range autocorrelation will produce large excursions of the cusum plot path, and the sampling of successive values close to their predecessors has the effect of smoothing the path. By contrast, a sequence of independent observations will be more irregular, or “hairy”. Representing on the same plot the cusum paths of different samplers run on the same data, like in section 3.8, is a good way to compare their mixing speeds. With a single sampler, Yu and Mykland [65] suggest plotting the cusum path of independent observations of the statistic as a benchmark. Judging whether mixing is satisfactory is more difficult in this case because the ideal mixing of an independent sequence cannot be achieved with a Markov chain. In this work, sequences of independent observations are obtained by permuting the order of the observations from a MCMC run.

When a large number of statistics are monitored in each MCMC run and a large number of analyses are performed, it becomes impractical to examine a cusum plot for each statistic. We then summarize the information provided by the cusum plot in a single number, the quantitative measure of the “hairiness” of a cusum path proposed by Brooks [3]. His argument is that a perfectly hairy path is one where increasing and decreasing segments alternate in successive intervals, so that every point is a local minimum or maximum, while a perfectly smooth path is a flat line with no extremum. Brooks [3] therefore measures hairiness as the proportion of points that are local minima or maxima. Letting

\[ d_i = \begin{cases} 1 & \text{if } U_{i-1} > U_i < U_{i+1} \text{or } U_{i-1} < U_i > U_{i+1} \\ 0 & \text{else,} \end{cases} \]

Brooks [3] defines \( D_t = \sum_{i=1}^{t-1} d_i/(t - 1) \). For an independent sequence of observations from a symmetric distribution, \( E[D_t] = \frac{1}{2} \), providing a point of reference.
Convergence of the empirical mean of a statistic to its expectation is measured by its variance. We state theoretical results on convergence of the MCMC estimate $\hat{\mu}_N = \sum_{r=1}^{N} g(S^{(r)})/N$ to the expectation $\mu = E[g(S)]$ for a general function of interest denoted $g$. The results are used to justify the estimators presented next.

The variance of $\hat{\mu}_N$ exists under the condition that $E[g^2(S)] < \infty$. Let $\gamma_t = \gamma_{t-} = Cov[g(S^{(r)}), g(S^{(r+t)})]$ be the lag $t$ autocovariance of $g$.

**Theorem 1** For a stationary, irreducible, reversible Markov chain, Kipnis and Varadhan [29] established that

$$N \text{Var}[\hat{\mu}_N] \to \sigma^2 = \sum_{t=-\infty}^{\infty} \gamma_t$$

almost surely. Provided that $\sigma^2 < \infty$, we have the following central limit theorem for $\hat{\mu}$

$$\sqrt{N}(\hat{\mu}_N - \mu) \overset{\mathcal{D}}{\to} N(0, \sigma^2)$$

and the stronger result

$$\sqrt{N}\frac{(\hat{\mu}_N[|N]) - \mu}{\sigma} \overset{\mathcal{D}}{\to} W_t$$

where $W_t$ is a Brownian motion.

The last result implies asymptotic independence of the means of subchains. For practical application an estimate of $\sigma^2$ is needed. Two types of estimators from Geyer [21] are applied.

**Batch means estimator** The N replicates are divided into $m$ batches of equal size $h = N/m$ and the mean of $g(S)$ computed for each batch.

$$\bar{X}_{N,k} = \frac{1}{h} \sum_{r=(k-1)h+1}^{kh} g(S^{(r)}), k = 1, \ldots, m$$

From theorem 1, the $\bar{X}_{N,k}$ tend to independent identically distributed random variables as $N$ goes to infinity. Geyer [21] comments that $N$ has to be much larger than the mixing time of the chain for the result to be valid, and with a finite length sequence $m$
must be kept small. A typical choice would be \( m = 10 \). The batch means estimator of the variance is based on the sample variance of the
\( \bar{X}_{N,k} \):

\[
\hat{\sigma}_B^2 = h s^2_{\bar{X}_N}
\]

Notice that this estimator is equal to the jackknife estimator of variance computed on pseudo-values obtained by deleting blocks of \( h \) observations. The main advantage of the batch means variance estimator is his simplicity. It does not require the entire sequence of realizations of \( g(S) \) but only the \( m \) batch means. With small \( m \) the estimate is however unstable. With \( m \) fixed it is also inconsistent, converging to a \( \chi^2_{m-1} \) distribution. The common practice in this case is to construct confidence intervals for \( \mu \) using the \( t \) distribution with \( m - 1 \) degrees of freedom, and those tend to be too wide. Also, the mixing properties of the chain have to be studied with another method on the entire sequence of observations for at least some of the summary statistics of a chain before one can be confident that the batches are long enough.

**Window estimators** There is more information in the full sequence of observations of a function of the states of a Markov chain than in the mean of the batches of those observations, and utilizing that information leads to more efficient estimators. The information from the sequence of observations is summarized in an estimate of the autocovariance function of the statistic.

\[
\hat{\gamma}_{N,t} = \frac{1}{N} \sum_{r=1}^{N-t} (g(S^{(r)}) - \bar{\mu}_N)(g(S^{(r+t)}) - \bar{\mu}_N)
\]

It might be tempting to estimate \( \sigma^2 \) by the sum of the estimates of autocovariance terms \( \sum_{t=-(N-1)}^{N-1} \hat{\gamma}_{N,t} \), but this estimator is inconsistent, since the number of autocovariance terms in the summation goes to infinity as fast as the autocovariance estimates converge. This problem can be remedied by downweighting terms as the lag increases, reducing the weight to 0 outside a certain window. The general form of this window estimate is given by

\[
\hat{\sigma}^2 = \sum_{t=-(N-1)}^{N-1} w_N(t) \hat{\gamma}_{N,t}, \quad 0 \leq w_N(t) \leq 1
\]

A weight function \( w_N(t) \) is defined by its shape and the width of the window where \( w_N(t) \) is greater than 0.
An extension of the jackknife estimator proposed by Künsch [33] exploits the full sequence of observations to estimate the variance. This intuitive estimator can be re-expressed as a window estimator with a particular window shape. The idea is to consider overlapping batches of $h = N/m$ observations. From the full sequence of observations, one can define $N - h + 1$ batches shifted by one observation. Redefining $\bar{X}_{N,j}$ as

$$\bar{X}_{N,j} = \frac{1}{h} \sum_{r=j}^{j+h} g(S^{(r)}), j = 1, \ldots, N - h$$

the jackknife estimator is given by

$$\hat{\sigma}_\text{Jack}^2 = \frac{N - h}{Nh} \sum_{j=0}^{N-h} (\bar{X}_{N,j} - \bar{X}_N)^2$$

where $\bar{X}_N = \sum_{j=0}^{N-h} \bar{X}_{N,j}/(N - h)$.

Künsch [33] showed that for estimating the variance of the sample mean this estimator is the same as the window estimator with weight function

$$w_N(t) = \begin{cases} 
1 - \frac{|t|}{h} & 0 \leq |t| \leq h \\
0 & |t| > h
\end{cases}$$

This weight function is known as the Bartlett window for spectral estimation (Priestley [45]).

The window width or bandwidth that minimizes the mean square error of the variance estimate (or other criteria to decide the bias-variance trade-off) depends on mixing properties of the Markov chain that are unknown. Geyer [21] proposed an adaptive bandwidth selection method based on properties of the autocovariance function of a Markov chain. He cites the following theorem:

**Theorem 2** For a stationary, irreducible, reversible Markov chain with autocovariance $\gamma_t$, let $\Gamma_m = \gamma_{2m} + \gamma_{2m+1}$ be the sum of consecutive pairs of autocovariances. Then $\Gamma_m$ is a strictly positive, strictly decreasing, strictly convex function of $m$.

Theorem 2 suggests criteria for selecting the bandwidth $h$ based on the properties of the theoretical autocovariance function. Using $\hat{\Gamma}_m = \hat{\gamma}_{2m} + \hat{\gamma}_{2m+1}$ as an estimate of $\Gamma_m$, those criteria are, from the less to the more stringent:
Initial positive sequence \( h = 2 \max \{ m : \hat{\Gamma}_r > 0, r = 1, \ldots, m \} + 1 \)

Initial monotone sequence \( h = 2 \max \{ m : \hat{\Gamma}_r > \hat{\Gamma}_{r-1}, r = 1, \ldots, m \} + 1 \)

Initial convex sequence \( h = 2 \max \{ m : \hat{\Gamma}_r \text{ convex minorant of } \hat{\Gamma}_1, \ldots, \hat{\Gamma}_r, r = 1, \ldots, m \} + 1 \)

With weights \( w_N(t) = 1, t = 1, \ldots, h \), Geyer [21] shows that the resulting estimate \( \hat{\sigma}^2 \) is almost surely greater than \( \sigma^2 \) for \( N > N_0 \) for the three definitions of \( h \). Without a way of determining \( N_0 \) this theorem offers no guarantee that the estimate from a finite sequence will be an overestimate, and Raftery and Lewis [47] present a counter-example where those estimators drastically underestimate the variance. Furthermore, \( \hat{\sigma}^2 \) is a truncated periodogram spectral estimator, a type of estimator sensitive to the value of \( h \) (Raftery and Lewis [47]). Given that the above criteria tend to give unstable estimates of the bandwidth, reducing sensitivity to the bandwidth is desirable, and the Bartlett weight function contributes to that.

Experience shows that, for the run lengths considered, small bumps in the estimated autocovariance function can prematurely terminate the initial monotone sequence and the initial convex sequence of \( \hat{\Gamma}_r \) while the general shape of the autocovariance function is still convex decreasing well beyond that point. Selecting the bandwidth using these criteria then leads to underestimation of the variance. The initial positive sequence in these cases is much longer and we observe in general the first negative \( \hat{\Gamma}_r \) only when the autocovariance function estimate is down to the noise level. We therefore use the initial positive sequence estimator. The decreasing weight of the autocovariance terms as the lag increases coming from the triangular shape of the window mitigates the potential overestimation.

3.8 Evaluation of convergence of different sampler combinations

An hybrid MCMC sampler can be constructed by taking any combination of the samplers described in sections 3.1 to 3.6. The relative proportions of different types of steps that optimize mixing depend on the pedigree analyzed and have to be studied empirically. In this section we compare different combinations of samplers for the accuracy and precision
of their estimates, their mixing behavior and their execution time on two test problems for which the computation of statistics of interest is pushing exact computing algorithms to their limits.

The locus sampler or the chain restarting sampler are included in all the sampler combinations because they generate an irreducible chain, making the chain generated by combining them with any other sampler irreducible. Since the locus sampler and chain restarting samplers tend to require much more computing time than the meiosis sampler or the parental sampler, we test combinations where the latter types of updates are performed in higher proportions. All computations were performed on a Sun Ultra5 workstation with a single 400 MHz processor.

The first test case is a 13 member pedigree on which gene segregation at 10 equidistant marker loci (10 cM intervals) with 4 equally frequent alleles and the locus of a dominant disease gene governing the disease phenotype located in the middle of the map has been simulated and the marker phenotypes retained for the 7 individuals in the last generation (figure 3.5). The dimensions of the meiosis indicator matrix for this case are $10 \times 18$. The conditional expectation of the statistic $S_{\text{pairs}}$ was computed using the enhanced version of the Lander-Green algorithm implemented in the Genehunter software package (Kruglyak et al. [32]) in 21 seconds. With the Lander-Green algorithm, memory space rather than computing time is the limiting factor, and adding two more non-founders to the test pedigree would have increased the size of the meiosis vector beyond Genehunter’s capabilities.

The second test case is a 72 member pedigree part of an actual larger pedigree of the Glaucoma Inheritance Study in Tasmania (figure 3.6). Phenotype data at 3 marker loci on chromosome 10 (D10S208, D10S220 and D10S1652) were sampled conditional on the original disease phenotypes recorded on the members of the pedigree, assuming a dominant gene responsible for the disease phenotype is located between D10S208 and D10S220. Only the individual phenotypes that were available in the actual study were retained. The meiosis indicator matrix has dimensions $3 \times 100$. The likelihood ratio between the hypotheses of a disease gene located at a series of 10 positions on the marker map and the hypothesis of a disease gene unlinked was computed under the trait model used to generate the genotypes at the trait locus using the pedigree peeling algorithm with set-recoding of the alleles.
implemented in the Vitesse software package (O’Connell and Weeks [42]). This 4-locus computation (3 markers plus the trait locus) required 26 minutes of computing time. With one more locus Vitesse exhausted all the memory space available on the machine.

The MCMC run length was set to 200,000 cycles of steps of any kind for all the sampler combinations considered. The chains were sampled every 20 cycles yielding 10,000 realizations in test case 1 and every 10 cycles yielding 20,000 realizations in test case 2. The sampling interval was set smaller in test case 2 because the variance was higher in that problem than in test case 1 but the gain in precision from denser sampling was higher because of the weaker autocorrelation. A burn-in of 2000 cycles in test case 1 and 1000 in test case 2 was discarded. Given the limited size of the test cases, the burn-in periods are

Figure 3.5: Pedigree and marker phenotypes for test case 1.
Figure 3.6: Pedigree used for test case 2. The individuals with observed marker phenotypes are marked with a +.

more than sufficient to reach the stationary distribution. Execution time varies considerably depending on the composition of the mix of samplers and was recorded for every run. In test case 1 the conditional expectation of $S_{pairs}$ was estimated and the bias computed at each of the 10 loci. The standard deviation (SD) of the estimates was computed at loci 1, 5 and 9 of the marker map where the full sequence of observations was recorded. Bias and SD were divided by the SD of the null distribution of $S_{pairs}$. The maximum standardized bias in absolute value over the 10 estimates is reported. In test case 2 the likelihood ratio (LR) was estimated at 10 locations at 2 cM intervals spanning the distance between D10S208 and D10S1652 by applying pedigree peeling for the trait locus conditional on the sampled meiosis indicators in a manner similar to the conditional peeling at a marker locus described in section 3.4, and averaging over all sampled realizations to get the estimate $\hat{LR}$ following the method of Sobel and Lange [52]. The relative bias $(\hat{LR} - LR)/LR$ was computed for the same locations. The entire sequence of realizations was saved at two locations, at -2 cM and 18 cM with respect to D10S208, to estimate the SD of $\hat{LR}$. We report the estimated SD divided by $\hat{LR}$. In both test cases the SDs were computed using the window estimator described in section 3.7. Cusum paths of the estimates from different combinations were plotted for one estimate. In test case 1 we picked locus 5, the locus where the variance of $S_{pairs}$ was highest. In test case 2 the cusum plot for the likelihood ratio at location 18cM was drawn.
The MCMC sampler implemented by Sobel and Lange [52] in the software package SIMWALK2 is also an hybrid of different types of steps, referred to as transition rules. Transition rules $T_0$ and $T_1$ are single locus versions of the meiosis and parental samplers respectively. Runs of the SIMWALK2 Markov chain were performed on the same test cases. The number of sampled pedigrees was set to 5000 in test case 1 and 10,000 in test case 2 to give execution times in the same range as the hybrid samplers in our implementation. All other parameters, including those determining the proportion of the different steps, were set to their default values. For test case 1, the allele sharing statistic $D$ was rescaled to estimate the conditional expectation of $S_{pairs}$ (not to be confounded with Brooks' mixing statistic $D_t$). Since the SIMWALK2 software does not output the values of the statistics for every sampled pedigree, the chain was stopped every 1000 sampled pedigrees to record the estimates up to those points. The intermediate results were used to derive the mean of every batch of 1000 realizations, and the SD of the batch means used as an estimate of standard deviation as described in section 3.7.

Tables 3.5 and 3.6 present the different combinations of samplers tested and the observed bias, SDs, and execution time of a MCMC run. Figures 3.7 to 3.9 are cusum plots of the estimates in test case 1 at locus 5 for various combinations of samplers in table 3.5 and figure 3.10 is a cusum plot of the estimates in test case 2 at location 18cM for combinations of samplers in table 3.6.

In test case 1, combining different types of samplers tends to reduce the bias and SD of the estimates, with the exception of the locus sampler alone at locus 1 with a lower SD than that of many hybrid samplers. Shorter excursions and more irregular cusum paths on figure 3.9 also indicate better mixing when more samplers are combined for locus 5. That trend is not present in test case 2; similar biases and standard errors are seen from the stand-alone locus sampler to the four sampler combination. One reason for this is the low autocorrelation between the realizations for all tested sampler combinations, indicating that they produce nearly independent observations. As a result of the high positive skewness of the distribution of the likelihood ratio, meiosis indicator matrices with very large LR values are sampled with low frequency. The combination of low correlation and high positive skewness produces cusum paths with big jumps followed by long irregular stretches exhibiting a
### Table 3.5: Convergence of $\hat{Z}$ with $Z = S_{pairs}$ for different combinations of MCMC samplers in test case 1. L = locus, M = meiosis, P = parental, R = chain restart.

| Run # | Step proportions | Max $|\text{std. bias}| \times 100$ | SD $\times 100$ | Execution time (min.) |
|-------|-------------------|--------------------|-----------------|----------------------|
|       |                   |                    | Locus 1        | Locus 5 | Locus 9 |                |
| 1     | L                 | 5.5                | 3.3            | 32.6    | 19.1    | 65              |
| 2a    | L:M               | 4.6                | 4.9            | 40.3    | 19.0    | 28              |
| 2b    | 1:5               | 1.0                | 4.7            | 43.3    | 22.6    | 33              |
| 2c    | 1:1               | 3.8                | 3.9            | 21.7    | 14.6    | 56              |
| 3a    | L:M:P             | 0.93               | 7.6            | 23.7    | 11.2    | 13              |
| 3b    | 1:5:5             | 0.90               | 6.6            | 19.4    | 10.5    | 25              |
| 3c    | 1:1:1             | 0.41               | 3.6            | 20.4    | 8.6     | 31              |
| 4a    | R:M:P             | 1.58               | 5.6            | 21.4    | 10.4    | 15              |
| 4c    | 1:1:1             | 0.91               | 3.7            | 9.4     | 6.8     | 42              |
| 5a    | L:R:M:P           | 1.24               | 5.1            | 16.8    | 9.6     | 17              |
| 5c    | 1:1:1:1           | 0.48               | 3.3            | 8.7     | 6.5     | 50              |
| 6     | SIMWALK2          | 1.12               | 2.9            | 4.7     | 11.6    | 38              |

### Table 3.6: Convergence of likelihood ratio with different combinations of MCMC samplers in test case 2. L = locus, M = meiosis, P = parental, R = chain restart.

| Run # | Step proportions | Max $|\text{rel. bias}| \times 100$ | Rel. SD $\times 100$ | Execution time (min.) |
|-------|-------------------|--------------------|----------------------|----------------------|
|       |                   |                    | -2cM 18 cM           |                      |
| 1     | L                 | 40                 | 5.4 30               | 1778                 |
| 2a    | L:M               | 34                 | 17 20               | 354                  |
| 2b    | 1:1               | 40                 | 13 28               | 900                  |
| 3a    | L:M:P             | 30                 | 5.3 23              | 156                  |
| 3b    | 1:1:1             | 31                 | 15 25               | 604                  |
| 4a    | R:M:P             | 34                 | 10 28               | 287                  |
| 4b    | 1:1:1             | 28                 | 19 28               | 1010                 |
| 5a    | L:R:M:P           | 35                 | 5.6 29              | 272                  |
| 5b    | 1:1:1:1           | 26                 | 14 28               | 1192                 |
| 6     | SIMWALK2          | 47                 | 13 31               | 153                  |
Figure 3.7: Cusum plot of $\hat{Z}$ with $Z = S_{pairs}$ for different proportions of locus and meiosis samplers in test case 1 at locus 5.

A downward trend in figure 3.10. The origin of the high skewness of the LR distribution is in the sampling of meiosis indicators ignoring the disease phenotype, as if the marker loci were unlinked to the disease locus. When a disease locus is actually present in the marker map, sampling assuming no linkage yields few meiosis indicator matrices consistent with linkage to the disease locus and those have large likelihood ratios (Thompson [59], chapter 9).

The lack of variance and bias improvement when the chain restarting sampler is added in test case 2 can be explained by the very low acceptance rate of the proposal, on the order of 0.001, compared to test case 1 where it ranged between 0.01 and 0.05. Test case 2 is at the limit of the number of meioses for which the acceptance probability of the chain restarting sampler is high enough for the proposed state to be accepted at least once in a realistic run length.
In test case 1, hybrid samplers with equal proportions of the included types of steps tend to give smaller bias and variance than samplers where the step proportions are skewed toward more meiosis and parental updates, and, at least for the locus-meiosis sampler, their cusum path at locus 5 has smaller excursions. A different behavior is observed in test case 2 where the SD at the location -2 cM is often lower with a high proportion of meiosis and parental sampler, but more or less unchanged at location 18 cM, preventing us from drawing general conclusions. This illustrates that the composition of the hybrid sampler affects the SD differently at different locations, a feature also observed in test case 1.

The execution times of the runs for corresponding combinations of steps are 10 to 30 times longer in the second than in the first test case. That increase is due in part to the larger size of the meiosis indicator matrix, but also to the presence of more allelic types in the test pedigree 2 that increases the computational cost of locus and chain restarting.
Figure 3.9: Cusum plot of $\hat{Z}$ with $Z = S_{pairs}$ for the locus, locus + meiosis, locus + meiosis + parental and locus + chain restart + meiosis + parental samplers in test case 1 at locus 5.

updates. The high cost of those updates is revealed by the increasing computing time as their proportion increases, and the increase is relatively larger on test pedigree 2. For a fixed computing time, a sampler with a low proportion of locus and/or chain restarting updates performs a larger number of cycles. The gain in accuracy and precision from the increase in sample size in most cases offsets the loss from using sub-optimal step proportions.

When compared to the best combination of samplers in each of the two test cases, the estimated SD of the estimates from the hybrid sampler of SIMWALK2 is lower for some statistics and higher for others, reinforcing the observation that a different construction of the Markov chain results in different relative levels of precision at different locations. The maximum absolute bias of the SIMWALK2 estimates is higher than the level achieved by the best combination of samplers but still in the range of other combinations tested. The
Figure 3.10: Cusum plot of likelihood ratio for the locus, locus + meiosis, locus + meiosis + parental and locus + chain restart + meiosis + parental samplers in test case 2 at location 18cM.

The ratio of the execution times of SIMWALK2 in test case 2 over test case 1 is smaller than for our hybrid samplers, due to the fact that none of the SIMWALK2 transitions requires to perform pedigree peeling.
Chapter 4

Searching for a gene predisposing to glaucoma in a pedigree from Tasmania

This chapter covers various aspects of an IBD sharing linkage analysis of glaucoma using the data from a genome scan on a large pedigree from Tasmania. The first section describes the form of glaucoma which is the disease phenotype of interest. The second presents the Glaucoma Inheritance Study in Tasmania, a large pedigree with multiple cases of glaucoma recruited as part of that study and the data available on that pedigree for linkage analysis. The hypotheses about the genes involved in the etiology of glaucoma in that pedigree stated in section 4.2.2 motivate the IBD sharing statistic $S_{most}$ and its extension introduced in section 4.3. Results from a simulation study of the power of those statistics and the standard statistic $S_{pairs}$ are then presented. The effect of varying the number of markers used to estimate the conditional expectation of IBD sharing statistics near a disease locus is also studied.

The MCMC methods presented in chapter 3 were applied to sample meiosis indicators in the pedigree conditional on all observed marker data on a chromosome for the 22 autosomes. The mixing behavior of the chains is investigated using the diagnostics presented at section 3.7 and estimates of the SD of IBD sharing statistic estimates are reported. Results of the estimation of the largest set of individuals sharing a gene IBD near a locus
where the actual largest set is known are reported next, followed by estimates of conditional expectations of statistics for the entire genome scan. The chapter concludes with a follow-up of the region with the highest estimates in the first MCMC run and attempts to reconcile estimates that are divergent due to poor mixing.

4.1 Primary open-angle glaucoma

Glaucoma is the general designation of an eye-condition characterized by optical nerve damage, visual field loss and, in a majority of cases, by an increase in intra-ocular pressure. Those symptoms are thought to be caused by a decreased rate of aqueous outflow through the trabecular meshwork, located in the iridio-corneal angle of the eye. Primary open-angle glaucoma (POAG) is the most common form of glaucoma. Open angle means that the aqueous humor still drains through the angle, but the drainage occurs at a reduced rate. The progressive decrease of the visual field resulting from POAG often leads to blindness. Glaucoma is the second leading cause of blindness worldwide after cataract.

POAG varies in severity and age of onset. The severe, juvenile onset form is rare while the adult-onset form accounts for the vast majority of POAG cases. Estimates of the prevalence of POAG in adults 40 years and older range from 1% to 6% depending on the ethnic group (Tielsch et al. [61]). A projection of prevalence estimates by age and ethnicity to the World population produces an estimated total number of cases of 33 millions (Quigley [46]).

Evidence of the role that genes play in the etiology of POAG comes from an increased risk in relatives of affected individuals compared to the population prevalence (Tielsch et al. [61], Wolfs et al. [64]) and the existence of families where POAG is inherited as a dominant trait. Genetic linkage analysis in families where the severe, juvenile onset form of POAG is inherited as a dominant trait led to the localization of the GLC1A locus on the q arm of chromosome 1 and later to the identification of the responsible gene, named trabecular meshwork-induced glucocorticoid response gene (TIGR) (Stone et al. [53]). Another group had independently identified the protein and called it myocilin (MYOC), the name later officially adopted. To date, 26 mutations in this gene have been found in glaucoma cases (Johnson [28]), different mutations being associated with different degrees of
severity of the disease. The role of the myocilin protein in the eye is currently the object of intense research. Mutations in the myocilin gene are found in only 3 to 5% of patients with adult onset POAG (Stone et al. [53], Fingert et al. [17]), and there are families with high concentration of POAG cases where no mutation in myocilin is found, indicating that other genes are involved. Linkage analysis in some of those families detected linkage to five other loci named GLC1B to GLC1F but the genes at those loci have not yet been identified, and it is possible that some of them are false positive results.

4.2 The pedigree GTAS2

Some characteristics of the population of Tasmania are advantageous for genetic studies. It was founded by a small group of settlers at the beginning of the 19th century and immigration to the island has been limited since then due to its geographical isolation. Accurate genealogical records have been preserved, making possible the reconstruction of pedigrees. The quality and universality of the health system facilitates the identification of disease cases in the population.

The Glaucoma Inheritance Study in Tasmania (GIST) (Sack et al. [49]) attempts to capitalize on those advantages to identify genes responsible for glaucoma. Its approach is to recruit pedigrees in the population of Tasmania for linkage analysis. The first step of the study was a census of all glaucoma cases in Tasmania, which identified 3000 cases. About 2000 of those were enrolled into the study. The research of the ancestry of those individuals led to the reconstruction of a number of pedigrees.

The largest of those pedigrees, with 2000 members, traces the ancestry of a group of glaucoma cases back to a founder couple six generations ago. This pedigree is identified as GTAS2, GTAS standing for Glaucoma in Tasmania. The pedigree contains 28 living individuals that were diagnosed with glaucoma following an eye examination. Except for one case of secondary glaucoma, all the affected individuals have adult-onset POAG with elevated intra-ocular pressure, the phenotype considered in the linkage analysis. Secondary glaucoma is a different form of the disease and the only case was excluded. An affected individual who married in the pedigree and her two affected children were also excluded to focus the analysis on affected individuals who are most likely to have inherited a disease
allele from an ancestor common to a large number of affecteds. The set of pedigree members assigned the phenotype “affected” for the linkage analysis therefore includes 24 individuals. The age at diagnosis of these individuals covers a wide range from 32 to 92 years, with a mean of 59 years and a SD of 16 years. While in young adults the diagnosis was made shortly after the onset of the disease, elderly people may have suffered from glaucoma for many years before being examined for the GIST, so the distribution of age at diagnosis is not reflecting the distribution of age of onset. To quantify the evidence for heritability of POAG in the pedigree, we obtained a crude estimate of the risk to siblings of 23% by identifying nuclear families with at least one case and counting the affected siblings of one index case per family (11) divided by the total number of index cases’ siblings (49). This rough estimate with no adjustment for age and environmental risk factors is close to the 22% lifetime risk estimate in siblings of POAG cases reported by Wolfs et al. [64]. The presence of 6 parents with affected children is another indication of the heritability of POAG in the pedigree GTAS2.

Screening for mutations in the myocilin gene revealed that 9 affected individuals had an identical mutation changing a codon coding for the amino acid glutamine into a stop codon, interrupting prematurely the polypeptide chain of the protein. Such a mutation is designated STOP mutation. That same STOP mutation is also present in 10 individuals manifesting no symptom of glaucoma. The mutation probably plays a role in the etiology of POAG in the carriers, but its marginal estimated penetrance of $9/19 = 0.47$ is low and the presence of 15 affected individuals without the mutation is evidence that other genes are involved.

Investigators obtained blood samples from 155 individuals, including all the affecteds. Those individuals and the individuals connecting them form a 259 member pedigree. After excluding the branch of the pedigree below the affected married-in individual and the nuclear family of the individual with secondary glaucoma we end up with the 246 member pedigree used for linkage analysis. This pedigree has 72 founders and 71 nuclear families. A drawing of the pedigree is presented in figure 4.1. Most affected individuals are found in the fourth and fifth generations; only one belongs to the most recent generation. Individuals in the last generation are in general too young to manifest symptoms of glaucoma. The advanced age of the affecteds has for consequence that few of their parents are still alive to
have their marker phenotypes determined.

4.2.1 Available marker data

The pedigree drawing shows that no marker data are available for all the individuals in the top three generations as well as in a majority of individuals in the fourth generation from the top, most of those people being deceased. In addition, marker data are missing on most of the living individuals who married in at the fourth and fifth generations because blood samples have not been collected from them. Not collecting marker data on the married-in individuals is a weakness of the study design, since determining the genotype of those individuals is important to establish which parental gene was transmitted to any given child, and from that information which genes are IBD between affected individuals. Assays to determine the marker phenotypes of a total of 127 individuals in the last three generations of the pedigree have been done.

A set of 401 short tandem repeat (STR) genetic markers covering the autosomal genome was selected for the linkage study. The genetic map of those markers was extracted from the Genetic Location Database [43]. Locus order and genetic distances in that database are estimates derived from the marker data collected on three-generation pedigrees from the CEPH (Dib et al. [11]). The map of the 22 chromosomes is shown on figure 4.2. The average estimated distance between consecutive markers is 8.5 cM and the maximum gap is 27.0 cM. STRs are highly polymorphic; the number of different alleles observed in the marker phenotypes of members of the pedigree ranges from 2 to 18 with a mean of 9.2. A plot of the distribution of the number of allele types at the 401 markers appears on figure 4.3. The allele frequencies of all the markers were estimated using a group of 72 individuals from Tasmanian nursing homes. The individuals in the group are of the same Caucasian ethnic background as the pedigree members. Even though the individuals do not form a random sample from the population, one can hope that the frequencies of alleles of markers with no biological function will not be correlated with the characteristics of the group. The SD of a frequency estimate of 0.10 from a sample of 144 chromosomes assumed to be independent is 0.025, a relatively large value. Allele frequencies are often found to have little influence on linkage statistics as long as they are not very small.
Figure 4.1: Pedigree GTAS2. Filled symbols represent individuals affected with POAG, shaded symbols represent individuals with no marker data and empty symbols represent individuals with observed marker phenotypes and no glaucoma symptoms.
Failures of the assay or inferred genotypes inconsistent with the rules of mendelian inheritance result in missing data. The percentage of missing phenotypes at a marker locus for the individuals who gave a blood sample ranged from 1% to 93% with a mean of 19%. Among the observed phenotypes, those that indicate homozygous genotypes range in proportion from 3% to 70% depending on the locus, with a mean of 23%. As a result, many of the markers provide very limited information on gene transmission when taken individually.

A likelihood analysis under a dominant genetic model has been conducted using one genetic marker at a time prior to our involvement in the study. The likelihood ratio statistic, or \textit{lod-score} as it is called in genetics, did not exceed established thresholds for genomewide statistical significance with any of the markers. Follow-up of regions showing a modest signal with additional markers did not provide evidence of the presence of a disease gene. The incompleteness of the marker data is the most likely reason why no strong signal was detected. A simple dominant model may also be inappropriate.

\section*{4.2.2 Hypotheses on the genetic causes of glaucoma in the affected individuals of the pedigree GTAS2}

Prior knowledge on the genetics of primary open-angle glaucoma, the structure of the pedigree and the distribution of the cases within the pedigree guided the formulation of hypotheses on the genetic causes of glaucoma in the affected individuals of the pedigree GTAS2. We formulate these hypotheses here and then introduce IBD sharing statistics inspired by those hypotheses in the following section.

The structure of the pedigree GTAS2 where all affected individuals are descendants from a common founder couple (figure 4.1) suggests that a disease allele inducing a strong predisposition to POAG may have been introduced by a top founder and passed on to a majority of the affecteds. An affected individual could have received only one copy of this allele, which would imply that the allele has a dominant effect. A dominant model with incomplete penetrance fits well the phenotype data in many glaucoma pedigrees (Sarfarazi \cite{50}) and the mutations found in the myocilin gene in families with cases of juvenile onset POAG have a dominant effect with incomplete penetrance (Stone \textit{et al.} \cite{53}). Cases of
<table>
<thead>
<tr>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
<th>Value 8</th>
<th>Value 9</th>
<th>Value 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.35</td>
<td>8.84</td>
<td>3.87</td>
<td>10.85</td>
<td>10.66</td>
<td>1.82</td>
<td>9.89</td>
<td>10.89</td>
<td>6.44</td>
<td>11.03</td>
</tr>
<tr>
<td>10.89</td>
<td>10.14</td>
<td>6.54</td>
<td>9.27</td>
<td>10.36</td>
<td>10.91</td>
<td>10.11</td>
<td>5.63</td>
<td>6.15</td>
<td>13.23</td>
</tr>
<tr>
<td>11.98</td>
<td>2.44</td>
<td>8.21</td>
<td>1</td>
<td>2</td>
<td>12.75</td>
<td>11.74</td>
<td>6.97</td>
<td>5.38</td>
<td>15.36</td>
</tr>
<tr>
<td>13.08</td>
<td>10.78</td>
<td>7.36</td>
<td>16.58</td>
<td>9.09</td>
<td>8.11</td>
<td>10.00</td>
<td>11.34</td>
<td>12.31</td>
<td>17.78</td>
</tr>
<tr>
<td>26.96</td>
<td>10.45</td>
<td>8.64</td>
<td>10.06</td>
<td>8.52</td>
<td>7.96</td>
<td>5.51</td>
<td>10.03</td>
<td>9.19</td>
<td>5.92</td>
</tr>
<tr>
<td>3.45</td>
<td>11.35</td>
<td>9.60</td>
<td>6.90</td>
<td>8.15</td>
<td>13.40</td>
<td>2.75</td>
<td>10.05</td>
<td>6.57</td>
<td>16.00</td>
</tr>
<tr>
<td>12.10</td>
<td>6.60</td>
<td>7.08</td>
<td>2.47</td>
<td>14.45</td>
<td>0.09</td>
<td>5.55</td>
<td>4.98</td>
<td>10.04</td>
<td>11.22</td>
</tr>
</tbody>
</table>

The table contains various numerical values arranged in a grid format.
Figure 4.2: Genetic map of the markers used in the genome scan of the GIST. Only the numeric portion of the marker labels are shown. Some loci numbers are not shown to avoid overlaps in dense regions of the map.

Figure 4.3: Number of allele types seen in the phenotypes of individuals from the pedigree GTAS2 at 401 marker loci.
overdominance where individuals homozygous for the disease allele are unaffected have also been reported (Morissette et al. [41]). In GTAS2 we find 6 instances where a parent and one or more of his children are affecteds, an observation consistent with a dominant mode of inheritance. However, there are also 5 affected individuals whose parents never manifested severe symptoms of glaucoma, an indication that, like for other complex traits, the effect of an hypothetical disease allele may be modified by variants in other genes and environmental factors. A single disease gene with a major dominant effect possibly modified by variants at other loci is the best case scenario to detect the location of a gene in the pedigree GTAS2.

Various forms of genetic heterogeneity may complicate the picture. Allelic heterogeneity means that two or more disease alleles of the same gene are introduced in the pedigree by the founders. This could happen in GTAS2 if the disease allele has high frequency in the population. Allelic heterogeneity is more likely between distant relatives whose genetic material comes from many founders other than their common ancestor(s). Figure 4.1 shows that in the pedigree GTAS2 a split occurs at the second generation. Affected individuals on opposite sides of the split between the descendents of the sisters 2007 and 2011 are more distantly related. If two distinct-by-descent disease alleles are present, a likely scenario would be that there is one segregating in each of the two main sub-pedigrees.

The role of a myocilin gene STOP mutation in the onset of POAG in 9 affected individuals is not established, but the presence of a mutation potentially contributing to glaucoma in a subset of affected individuals of the GTAS2 pedigree means that more than one gene could be involved, a phenomenon referred to as locus heterogeneity. A third scenario can therefore be formulated with two unknown genes, each with a high risk allele. The split in the pedigree suggests again the possibility that a different gene may be acting in each of the two sub-pedigrees. The distribution of the STOP mutation in the myocilin gene between the affecteds shows however that within a sub-pedigree only a subset of affecteds may be carrying a mutation.

4.3 Largest set of alleles IBD in affected individuals

This section defines an IBD sharing statistic to detect genes under the scenario where a single disease allele introduced by a founder confers a high predisposition to the
disease. Considering the IBD configuration of the affected individuals at the disease locus, we expect to find one large class of IBD genes (genes in the sense of pieces of genetic information in individuals, cf. section 2.2.2) formed by copies of an original disease allele whose carriers have a high probability of being among the group of affecteds. The size of that largest class is the function of the IBD configuration that we propose to use and that we call $S_{\text{most}}$ for “most represented gene” in the affecteds.

**Definition 1** $S_{\text{most}}$: cardinality of the largest class of genes IBD among affected individuals in a pedigree.

When no affected individual is inbred, each individual may carry at most one copy of the gene. An equivalent definition of $S_{\text{most}}$ is then:

**Definition 2** $S_{\text{most}}$ (no inbred affected individual): largest number of affecteds sharing a gene IBD.

The definitions make no assumption on the origin of the allele found in largest number in the affecteds. The largest set of IBD genes will however be close to its maximum size only if the gene originates from one of the top founders. The statistic should be applied only to sets of affecteds having one or more common ancestors. An *ad hoc* extension to a pedigree with multiple founder couples ancestral to overlapping subsets of affecteds is to take the sum of $S_{\text{most}}$ computed on each subset (Greenwood *et al.* [22]).

Sobel and Lange [52] arrived via a different route to the same statistic and called it statistic B. McPeek [39] defined a statistic called $S_{\text{rob dom}}$ for “robust dominant” that was derived from the likelihood ratio statistic under a two-allele quasi-dominant model, letting the disease allele frequency go to 0. Under that model, an individual carrying at least one copy of the $A$ allele has a risk $f_1$ of being affected $r$ times greater than the risk $f_0$ of an individual carrying two copies of the $a$ allele (like in section 2.2.1). After deriving $S_{\text{rob dom}}$ we show that it is equivalent to $S_{\text{most}}$ when the relative risk $r$ is large.

The IBD configuration $B$ of the $k$ affected individuals in a pedigree contains $J$ distinct by descent classes of genes labeled from 1 to $J$. Let $n(j)$ be the number of affected
individuals in the pedigree with at least one copy of a gene from class $j = 1, \ldots, J$ and $A$ be the subset of gene classes that are of allele type $A$. The genes ancestral to each of the $J$ classes come from founders of the pedigree. It is assumed that those founders were sampled independently from a population in Hardy-Weinberg equilibrium where allele $A$ is present with frequency $p$, so that the number of genes of allele type $A$ among the $J$ ancestral genes follows a binomial distribution with probability of success equal to $p$. The probability of disease phenotype $Y_D$ conditional on IBD configuration $B$ under the hypothesis $H_1$ that $B$ is observed at the disease locus can be expressed as a polynomial in $p$. In the following, $p$ is assumed small and terms of second order and higher are ignored.

$$P_{H_1}[Y_D|B] = P[Y_D, A = \emptyset|B] + \sum_{j=1}^{J} P[Y_D, A = \{j\}|B] + \ldots$$

$$= \prod_{i=1}^{k} P[Y_{D_i}|A = \emptyset, B] P[A = \emptyset]$$

$$+ \sum_{j=1}^{J} \prod_{i=1}^{k} P[Y_{D_i}|A = \{j\}, B] P[A = \{j\}] + \ldots$$

$$= f_0^k (1 - p)^J + \sum_{j=1}^{J} f_j^n (1 - p)^{k-n} p(1 - p)^{J-1} + o(p)$$

$$= f_0^k \left[ 1 - p \sum_{j=1}^{J} (f_j / f_0)^n (1 - 1) + o(p) \right]$$

$$= f_0^k \left[ (1 + p \sum_{j=1}^{J} (r^n - 1) \right] + o(p)$$

Now $P[Y_D] = \sum_B P_{H_1}[Y_D|B] P[B] = E_{H_0} [P_A[Y_D|B]]$. Therefore, we can write the likelihood ratio as

$$LR(Y_D, B) = \frac{P_{H_1}[Y_D|B]}{E_{H_0} [P_{H_1}[Y_D|B]]} = \frac{f_0^k \left[ (1 + p \sum_{j=1}^{J} (r^n - 1) \right] + o(p)}{f_0^k \left[ (1 + p E_0 \sum_{j=1}^{J} (r^n - 1)) \right] + o(p)}$$

$$= 1 + p \left[ \sum_{j=1}^{J} (r^n - 1) - E_{H_0} \left[ \sum_{j=1}^{J} (r^n - 1) \right] \right] + o(p)$$
When \( p \downarrow 0 \) the likelihood ratio is governed by the first order term. This term depends on the relative risk \( r \). McPeek [39] reports that “in practice, the power to detect linkage is not very sensitive to the choice of \( r \)” and arbitrarily chooses \( r = 7 \) in the definition of \( S_{rob\ dom} = \sum_{j=1}^{J} (7^{n(j)} - 1) \).

Provided \( r \) is large enough, say \( r \geq 5 \), the term in the summation for the gene \( j \) with the largest \( n(j) \), in the absence of a tie, will dominate since it is at least \( r \) times larger than the next largest (or \( r/2 \) times larger than the sum of two subsets tied for second largest). Then, keeping only the term with the largest \( n(j) \) gives a statistic almost equivalent to the sum over all \( j \). That statistic is a monotone function of \( \max n(j) \), which corresponds to definition 2 of \( S_{most} \). When there is a tie for the largest \( n(j) \) more than one term dominates the summation. This is most likely to happen at loci where there is no rare allele with dominant effect. Reducing the score of that outcome by half by picking one of the two subsets of equal size should therefore have no negative impact on the power of the statistic to detect dominant genes.

4.3.1 Extension to potentially heterogeneous sub-pedigrees

The statistic \( S_{most} \) is expected to have power to detect a disease locus when most affected individuals in a pedigree share a single disease allele IBD. Its power will be reduced if multiple distinct-by-descent disease predisposing alleles segregate in the pedigree, at the same or at different loci. Genetic heterogeneity within a pedigree complicates considerably the detection of a DS gene. The power of a linkage detection procedure may be maintained if sub-pedigrees in which different disease alleles may be present are pre-identified. We propose here an extension of the statistic \( S_{most} \) to the case of two potentially heterogeneous sub-pedigrees. This approach is motivated by the structure of the pedigree GTAS2 which splits into two main sub-pedigrees at the first generation below the founder couple (figure 4.1).

In that situation, the allele with the largest number of IBD copies in the set of affected individuals may either be found in affecteds in both sub-pedigrees, implying that it originates from a top founder, or in affecteds in only one of the two sub-pedigrees. In the latter case, we propose to compute the largest number of affecteds sharing IBD in each
sub-pedigree. We define statistics extending $S_{\text{most}}$ to sub-pedigrees:

\[ S_i = \text{Size of largest set of genes IBD in the affecteds of sub-pedigree } i \text{ sharing IBD} \]

\[ S_i^* = \text{Size of largest set of genes IBD in the affecteds of the whole pedigree that includes the largest set of genes IBD in the affecteds of sub-pedigree } i \]

The distribution of the affecteds sharing IBD among the two sub-pedigrees gives rise to three mutually exclusive events:

1. The largest class of genes IBD spans both sides of the pedigree and is the union of the largest class of genes IBD in each sub-pedigree, i.e. $S_1^* = S_2^* = S_1 + S_2$.

2. The largest class of genes IBD spans both sides of the pedigree but the largest class of genes IBD in one of the sub-pedigrees is not included in that set, i.e. $S_1 + S_2 > \max(S_1^*, S_2^*) > \max(S_1, S_2)$.

3. The largest class of genes IBD does not span both sides of the pedigree $\max(S_1, S_2) + \min(S_1^*, S_2^*) > \max(S_1^*, S_2^*) = \max(S_1, S_2)$.

As long as the largest class of genes IBD spans both sub-pedigrees, the statistic $S_{\text{most}} = \max(S_1^*, S_2^*)$ reflects the fact that under the hypothesis of a single major disease causing allele in the pedigree, it is likely to be present in affecteds in both sub-pedigrees. If it is not, then there may still be a disease allele confined to one of the sub-pedigrees or two distinct-by-descent disease alleles in the two sub-pedigrees. In that case we measure large classes of genes IBD in one or both of the sub-pedigrees.

Let \( I = S_{\text{most}} > \max(S_1, S_2) \). The event \( I = 1 \) means that the largest class of genes IBD spans both sides of the pedigree. When \( I = 0 \) and the two sub-pedigrees are considered separately, the statistics \( S_1 \) and \( S_2 \) have to be combined in some way. Defining a weighted sum of \( S_1 \) and \( S_2 \) is one possibility. Taking the sum of \( S_1 \) and \( S_2 \) on their original scale implies that each pedigree receives an importance roughly proportional to the number of affected individuals it contains. Weighting the statistic values by the inverse of the number of affecteds would give equal importance to the two pedigrees.

\( S_1 \) and \( S_2 \) are not independent under the null hypothesis of no disease gene linked even conditional on \( I = 0 \) because classes of genes smaller than the largest class may overlap
the two sub-pedigrees. It is therefore preferable to estimate the null distribution of \( S_1 + S_2 \) by sampling from the joint distribution of \( S_1 \) and \( S_2 \) conditional on \( I = 0 \) and computing \( S_1 + S_2 \) for each realization.

The dependence between \( S_1 \) and \( S_2 \) can be expected to be weak given that the subsets of genes are constrained to be in different sub-pedigrees. If the dependency is ignored, an alternative significance assessment procedure is to compute the p-value \( P_1 \) of \( S_1 \) and \( P_2 \) of \( S_2 \) from their respective marginal distributions and to use the product \( P_1 P_2 \) as combined statistic. That scheme gives the two sub-pedigrees the same weight. Under the null hypothesis \( P_1 \) and \( P_2 \) are uniformly distributed, and the monotone function \( P_1 P_2 \cdot (1 - \log(P_1 P_2)) \) of the product \( P_1 P_2 \) is also uniformly distributed, providing a p-value for the statistic.

### 4.4 Power of IBD sharing statistics

The power to detect a disease gene using an IBD sharing statistic depends on three main factors:

1. The distribution of copies of disease alleles among affected relatives resulting from the effect of all genes and environmental factors involved in the etiology of the disease. That distribution is reflected in the patterns of IBD sharing between affected individuals at a disease gene locus.

2. The rate of decline from the peak IBD sharing at a disease gene locus to the background level of IBD sharing which depends on the number of meioses separating affected individuals.

3. The ability to infer patterns of IBD sharing from incomplete marker data, and the degree of completeness of those data.

In section 4.4.1, we study the first of these factors by simulation under the idealized situation where the IBD configuration is observed at a disease locus, while the third factor is investigated in section 4.4.2 with a simulation study of the estimation of IBD sharing statistics by their conditional expectation computed given the data from different numbers of genetic markers in a multilocus analysis.
4.4.1 Power of IBD sharing statistics under the hypotheses on genetic causes of glaucoma in the pedigree GTAS2

The power of IBD sharing statistics is evaluated in a simulation study on the structure of the pedigree GTAS2. For each of the scenarios described below, the type of the founder alleles was fixed and the transmission of those alleles to non-founders simulated under Mendel laws.

The hypotheses stated in section 4.2.2 do not specify penetrances of the genotypes that may vary depending on the effect of other unknown genes. For the simulations the penetrances were fixed to values in the range expected under the hypotheses. The model of the action of each gene taken individually is quasi-dominant. The two-gene model is an heterogeneity model, where either of the two genes confers a high risk of developing the disease. The penetrance values for one and two-gene models are given in table 4.1.

<table>
<thead>
<tr>
<th>Geno.</th>
<th>Penetrances</th>
</tr>
</thead>
<tbody>
<tr>
<td>One gene</td>
<td>0.05 0.90</td>
</tr>
<tr>
<td>Two genes</td>
<td>0.01 0.90</td>
</tr>
</tbody>
</table>

Table 4.1: Penetrances of the quasi-dominant one-gene and two-gene models for the power study on the pedigree GTAS2.

Disease phenotypes were generated according to the penetrances given the genotypes of the pedigree members. A replicate of genotype and phenotype data was retained if it met minimal conditions that would lead to the recruitment of the whole pedigree in a family study. Those conditions are the presence of at least 20 affected individuals and at least one affected individual in each of the two main sub-pedigrees formed by the descendants of 2007 and 2011.

The power of IBD sharing statistics was evaluated under the following scenarios for the introduction of disease alleles by the founders:

1. One-gene model; single disease allele introduced by one of the two top founders of the
pedigree GTAS2. This corresponds to hypothesis 1. (92 replicates.)

2. One-gene model; two disease alleles. One allele introduced by one of the top founders, the other by 2012, the other founder of the large sub-pedigree. This is consistent with hypothesis 2. (89 replicates.)

3. Two-gene model; disease allele of gene $A$ introduced by a top founder.

   (a) disease allele of gene $B$ introduced by 2010, another founder of the small sub-pedigree. (102 replicates.)

   (b) disease allele of gene $B$ introduced by 2012, another founder of the large sub-pedigree.

   (93 replicates.)

Those scenarios are consistent with hypothesis 3.

The statistics $S_{most}$, the extension of $S_{most}$ to two heterogeneous sub-pedigrees and $S_{pairs}$ as defined in section 2.2.3 were evaluated. For the extension of $S_{most}$, the unscaled statistics for each sub-pedigree were summed. The power was evaluated with a critical value for the statistics corresponding to a significance level of 0.05 in a continuous genome scan (see section 2.3.1). For two gene models, IBD sharing was measured at gene $A$. The results are presented in figures 4.4 and 4.5.

![Figure 4.4: Power of IBD sharing statistics under scenarios 1 and 2 on pedigree GTAS2.](image-url)
The power ranges between 75% and 85% for all three statistics in scenarios 1 to 3a. Under scenario 3b where two alleles in different genes are present in the largest sub-pedigree, in most replicates gene $A$ contributes too few cases to be detected by the IBD sharing statistics, and the power drops to between 20% and 30%. The power of $S_{pairs}$ equals or, under scenario 2, slightly exceeds the power of $S_{most}$ computed on the entire pedigree. Measuring joint sharing of genes by a large group of affecteds appears not to improve power compared to a measure of pairwise sharing. The version of $S_{most}$ extended to sub-pedigrees had higher power than the two other statistics in the case of allelic heterogeneity (scenario 2). This was expected since the statistic was designed for that hypothesis. Its power is comparable or slightly lower than the power of the other two statistics under the three other scenarios.

4.4.2 Estimation of IBD sharing statistics by their conditional expectation computed in a multilocus analysis using different numbers of markers

This section addresses the inference of IBD sharing from incomplete marker data. A simulation study was conducted to investigate the effect of including an increasing number of markers in the analysis on the conditional expectation of IBD sharing statistics near a locus where a large proportion of affecteds share a gene IBD.

We attempted to recreate as much as possible the conditions of a real genome scan...
in a large pedigree. We picked a 50 cM region of chromosome 10 and used the map locations and allele frequencies of 6 markers in that region from the genome scan of the GIST. The genetic map of that region is depicted in figure 4.6. The number of allele types of the 6 markers ranges from 7 to 13. A disease locus was then placed at the mid-point between D10S208 and D10S220.

![Genetic map of the region of chromosome 10 used for data simulation.](image)

The pedigree structure used in the simulation study is the same one as in the second test case of the study comparing different MCMC sampler combinations and is pictured in figure 3.6. We chose to use that pedigree instead of the larger GTAS2 pedigree because of our greater confidence in the convergence of MCMC estimates within an achievable computing time on that smaller pedigree. The selected pedigree exhibits a pattern of missing data on the ancestors typical of large multigeneration pedigrees.

One of the top founders was assigned a disease allele $A$. All other founder alleles are normal $a$. Replicates of disease locus genotypes were simulated by dropping the genes from generation to generation, then disease phenotypes were generated under the disease model of table 4.2.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Penetrance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$aa$</td>
<td>0.02</td>
</tr>
<tr>
<td>$Aa$</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 4.2: Penetrances used to simulate disease phenotypes.

The marker genotypes of the pedigree founders were generated under the assumptions of Hardy-Weinberg and linkage equilibrium (see assumptions 2 and 4) using the allele frequencies estimated from the population of Tasmania. The segregation of the marker
alleles was generated conditional on the gene transmission at the disease locus assuming no crossover interference. The order of the alleles was ignored to output marker phenotypes. Marker phenotypes of individuals unavailable in the study were discarded, as were the marker phenotypes of available individuals at loci where the data were missing in the actual dataset.

We compare the values of $\bar{S}_{\text{pairs}}$ and $\bar{S}_{\text{most}}$ computed conditional on the data from one, three and all six marker loci to the mean values of the corresponding statistics. Estimation of $\bar{S}_{\text{pairs}}$ and $\bar{S}_{\text{most}}$ given marker data at three and six loci was done using MCMC sampling of meiosis indicators. The whole meiosis and whole locus Gibbs samplers as well as the parental Metropolis sampler were used. It was found in section 3.8 that combining those three samplers gave good precision on that pedigree. The proportions of cycles of those samplers was set to one locus sampler cycle for 10 meiosis sampler and 10 parental sampler cycles to limit computing time. Runs of 70,000 cycles were performed and the Markov chains were sampled every 7 cycles, producing a sample of 10,000 realizations of $\mathbf{S}$ for each run. The sampling interval was shortened compared to section 3.8 following the observation that the autocorrelation of the statistics is very low with that pedigree and number of markers. The variance of the estimates computed from the means of batches of realizations was small compared to the variability between replicates of the pedigree data and the trace of the cumulative mean of the statistics has stabilized after that run length (data not shown).

The expectations $\bar{S}_{\text{pairs}}$ and $\bar{S}_{\text{most}}$ conditioned on the marker data at a single locus could in principle be computed exactly using the pedigree peeling algorithm. SimIBD is the only publicly available program that we found able to compute $\bar{S}_{\text{pairs}}$ with a single marker on the test pedigree, but instead of performing an exact computation, it samples a number of replicates of the unobserved marker genotypes conditional on the observed ones, and for each realization of the complete genotypes (which may contain uninformative meioses) computes the IBD probability for all pairs of alleles carried by the affecteds. The locus sampler can also be used to draw independent realizations of meiosis indicators given the marker data at one locus and obtain a Monte Carlo estimate of $\bar{S}_{\text{pairs}}$ and $\bar{S}_{\text{most}}$. Based on the sample SD of $S_{\text{pairs}}$ and $S_{\text{most}}$ it was found that a sample size of 1000 is sufficient to reduce the SD of the sample average to a negligible level. The difference between the
estimates of $\tilde{S}_{pairs}$ from the locus sampler and from SimIBD were within the range expected by chance (data not shown.) Estimates from the locus sampler are reported here.

Figure 4.7 and 4.8 show the mean value over the 20 replicates of the IBD sharing statistics $S_{most}$ and $S_{pairs}$ respectively at each locus. The mean value of $\tilde{S}_{pairs}$ and $\tilde{S}_{most}$ at a locus given the marker data at one, three and six loci are presented on the same figures. Those mean values can be seen as estimates of the mean of the actual statistics and their conditional expectations given marker data at the selected loci in a population of pedigrees with the same disease allele type introduced by a founder at the same locus.

Figures 4.7: Mean of the actual values and conditional expectations of $S_{most}$ in 20 replicates of the phenotype data on the pedigree of figure 3.6. Conditional expectations are computed at the six loci given the data at the same six loci, at two groups of three consecutive loci given the data at the three loci in each group, and at the six loci given the data at each locus.

The conditional expectation of the statistics tends to be closer to the true statistic value as the number of markers used to compute $\tilde{S}_{pairs}$ and $\tilde{S}_{most}$ increases. There is a substantial improvement in the level of the statistics from one to three markers and a more modest one from three to six markers. The gap between the mean statistic value and the conditional expectation given data at six markers remains about as large as the
Figure 4.8: Mean of the actual values and conditional expectations of $S_{pairs}$ in 20 replicates of the phenotype data on the pedigree of figure 3.6. Conditional expectations are computed at the six loci given the data at the same six loci, at two groups of three consecutive loci given the data at the three loci in each group, and at the six loci given the data at each locus.

The estimated mean difference between conditional expectations computed using three and six markers and between six and one marker with simultaneous 95% confidence intervals for the two differences at each locus separately are shown on figure 4.9 for $S_{most}$ and figure 4.10 for $S_{pairs}$. With the small number of replicates that it was feasible to obtain, the confidence intervals remain quite large. Intervals for the difference between conditional expectations given data at six and three markers overlap 0 in most cases except at D10S208, a locus flanking the disease locus and where the difference is largest. The estimated difference between the conditional expectation given data at six and one marker loci are significantly less than 0 at the five loci closest to the disease locus. The conditional expectation of the statistics computed given the data at three marker loci are slightly higher for the trio D10S548, D10S197 and D10S208 than for D10S208, D10S220 and D10S1652 even though the disease locus is located in the interval between D10S208 and D10S220 but the differences are much smaller than the confidence intervals.

The large underestimation of the true value of IBD sharing statistics near a disease
Figure 4.9: Estimated mean difference between $\tilde{S}_{\text{most}}$ computed using six and three markers and difference between six and one marker at six loci on chromosome 10 for the pedigree of figure 3.6 with simultaneous 95% confidence intervals for the two differences at each locus separately.

Figure 4.10: Estimated mean difference between $\tilde{S}_{\text{most}}$ computed using six and three markers and difference between six and one marker at six loci on chromosome 10 for the pedigree of figure 3.6 with simultaneous 95% confidence intervals for the two differences at each locus separately.
locus where sharing is high translates into a major loss of power if the distribution of
the statistics assuming complete information is used as the null distribution for assessing
significance. The spacing of the markers as well as the proportion and relationships of the
individuals on which data is available is representative of actual linkage studies. This finding
is a strong incentive to attempt to approximate the null distribution of the conditional
expectation itself in order to obtain less conservative p-values and mitigate the power loss.

4.5 Parameters of the genome scan analysis

All the marker loci on a chromosome were analyzed jointly. For the first pass
on the whole genome a single MCMC run was performed for each chromosome to sample
the meiosis indicators at all the loci of a chromosome conditionally on marker data at all loci.

A combination of the whole meiosis and whole locus Gibbs samplers and the
parental Metropolis sampler was selected, following the conclusion of the study of sec-
tion 3.8 that combining multiple samplers improves precision and accuracy. Attempts to
use the chain restarting sampler revealed that the acceptance probability was too low for a
proposed state to be accepted even once in a run with several thousands trials of the chain
restarting step. The size of the pedigree GTAS2 is beyond the range where it is practical
to apply that sampler. One cycle of the locus sampler was performed for every 10 cycles of
meiosis and parental samplers. This choice was made because of the high amount of com-
puting time required by each locus sampler step. The larger number of cycles performed in
the same amount of time results in a gain in precision and accuracy that more than com-
pensates the potential inefficiency due to the unbalance in the sampler cycle proportions.
Realizations of meiosis indicators were sampled every 21 cycles, after the completion of a
round of all the samplers.

The Markov chain was run until a sample of 8000 realizations divided into batches
of 2000 had been generated. Further blocks of 2000 realizations were sampled depend-
ing on the magnitude of the fluctuations in the statistic estimates and the condensed IBD
configurations described in section 4.6.6 observed from batch to batch. The assessment of
convergence was visual and not based on a well defined criterion. No chain was run beyond
16,000 realizations. Sampling a batch of 2000 realizations takes between three hours of
computing time with 5 loci on chromosome 21 and 64 hours with 34 loci on chromosome 1 on a Sun Ultra5 workstation with 400 MHz processor.

The values of the IBD sharing statistics $S_{pairs}$ and $S_{most}$ at every locus were computed from each realization of $S$ for the whole set of 24 affected individuals and for the subset of 15 affecteds who do not carry a mutation in the myocilin gene. In a slight abuse of notation, we will denote by $\hat{S}_{most}$ and $\hat{S}_{pairs}$ the estimates of conditional expectations, not to be confused with the meiosis indicators $S_{li}$. The condensed IBD configurations at every locus described in section 4.6.6 were recorded for every realization. These IBD configurations were used among other things for computing the generalized version of $S_{most}$ allowing for heterogeneity between sub-pedigrees. A likelihood ratio statistic under a dominant model with a disease allele frequency of 0.001 was also computed.

4.6 Convergence monitoring for the analysis of the genome scan on the pedigree GTAS2

4.6.1 Statistics monitored

One form of convergence assessment was done using the IBD sharing statistics for which we estimate the conditional expectation given the marker data. The entire sequence of realizations of the IBD sharing statistics $S_{most}$ and $S_{pairs}$ for the whole set of 24 affected individuals was recorded at every fourth locus on the genetic map. Considering every fourth locus saves disk space while still providing multiple statistics to monitor in each run. The spacing of roughly 40 cM between the selected loci limits the correlation between statistics at different loci. Sequences of observations were recorded at 108 of the 401 marker loci on the genome. In addition, the mean of the statistics for both the full set of 24 and the reduced set of 15 affecteds was computed on batches of 2000 realizations at all 401 marker loci to compute the batch means estimate of the variance of $\hat{S}_{most}$ and $\hat{S}_{pairs}$.

4.6.2 Mixing diagnostics results

The $D$ statistic of Brooks [2] was computed on the entire sequence of realizations of $S_{most}$ and $S_{pairs}$ at each of the 108 loci where it was available and its values are shown on figure 4.11. We examined cusum plots of the estimates at loci with a very low value of
$D$ and loci with a value of $D$ near $\frac{1}{2}$ for both statistics to check that $D$ reflects the visually perceived quality of mixing.

![Figure 4.11: Statistic D for $S_{\text{most}}$ and $S_{\text{pairs}}$ for the set of 24 affected individuals in the pedigree GTAS2.](image)

The range of $D$ is wider for $S_{\text{most}}$ than $S_{\text{pairs}}$, some values of $D$ for $S_{\text{most}}$ being close to 0. The smaller number of distinct values of $S_{\text{most}}$ is an important factor explaining the lower number of local minima and maxima for $S_{\text{most}}$ compared to $S_{\text{pairs}}$. At the loci with $D$ close to 0, the low value is due to the concentration of the distribution of $S_{\text{most}}$ on one value more than to the mixing speed. For instance, at D5S2115 where $D$ is the lowest for $S_{\text{most}}$, 99.7% of the observations are equal.

Values of $D$ lower than 0.25 for either statistic are observed for most of the runs except the runs for chromosome 3, 11, 17 and 18. Very low and and very high values of $D$ are seen at different loci in the same run, indicating that the mixing speed for a meiosis
vector may differ widely from locus to locus within the same multilocus model.

We consider now loci where the distribution of the statistics is not concentrated on one value and $D$ is nevertheless low. The lowest value of $D$ for $S_{pairs}$ is observed at D6S422 (0.139), a locus with also a low value of $D$ for $S_{most}$ (0.062). The cusum plots for $S_{most}$ and $S_{pairs}$ at this locus are shown on figures 4.12 and 4.13. The two paths have remarkably similar shapes. The largest excursion of the two paths are more than 20 times larger than the largest excursion of the scrambled order and the paths are smooth. Successive stretches of realizations remaining within a small range followed by a shift to another range are apparent on the plot. All those are strong indications that the sampler takes a long time to move between states with different statistic values.

![Figure 4.12: Cusum plot for $S_{most}$ at D6S422.](image)

The locus where the value of $D$ is closest to $\frac{1}{2}$ is D2S396 (0.502 for $S_{most}$ and 0.499 for $S_{pairs}$). The cusum plots at that locus are presented in figures 4.14 and 4.15. The paths...
on those plots are much more irregular than at the previous locus, and aside from a longer excursion near the middle of the sequence, they cannot be distinguished from the path of the permuted sequence.

4.6.3 Standard deviation estimates

Estimates of the SD of $\hat{S}_{\text{most}}$ and $\hat{S}_{\text{pairs}}$ were computed using either the batch means estimator or the window estimator. It is important to keep in mind that the validity of those estimates rests on the assumption that the Markov chain is mixing fast enough that the interval between independent observations is shorter than the length of the sequence, a condition that may not be satisfied at many of the loci according to the diagnostics of the previous section. When mixing is poor, the SD may be underestimated.

The Bartlett window variance estimator with adaptive selection of the bandwidth is more efficient than the batch means estimator, but since its computation requires the full
sequence of realizations of the statistic, it could only be computed at every fourth locus, and only for $S_{\text{most}}$ and $S_{\text{pairs}}$ with 24 affected individuals. Here we compare the window estimator to the batch means estimator with batches of size 2000 at the loci where both estimates were computed.

Figures 4.16 and 4.17 show the log base 10 ratio of the batch means estimator over the window estimator of the SD of $\hat{S}_{\text{pairs}}$ and $\hat{S}_{\text{most}}$ plotted against the window estimator. For both statistics the ratio of the SD estimates falls within the band from $\frac{1}{2}$ to 2 for most loci. There are however exceptions where one of the estimates is up to ten times greater than the other. The alignment of points along decreasing curves on the plot for $S_{\text{most}}$ is an artifact of rounding off batch means SD estimates to the second decimal. This also explains the absence of points in the lower left corner of the plot. The overall good agreement between the two estimates indicates that the batch means estimator is a reasonable substitute when the window estimator is not available.
The maximum and the median of the window and batch means estimates of the SD are presented in table 4.3. Those SD estimates are small compared to the SD of the null distribution of the statistics which is 1.66 for \( S_{\text{most}} \) and 10.87 for \( S_{\text{pairs}} \). In the worst cases, the SD estimate is only 48% of the SD of the null distribution for \( S_{\text{most}} \) and 30% for \( S_{\text{pairs}} \), and for most loci the percentage is much lower.
Figure 4.16: Log base 10 of the ratio of the batch means estimator of the SD of $\hat{S}_{\text{pairs}}$ over the window estimator against the window estimator. Dotted lines drawn at $\frac{1}{2}$ and 2.

Figure 4.17: Log base 10 of the ratio of the batch means estimator of the SD of $\hat{S}_{\text{most}}$ over the window estimator against the window estimator. Dotted lines drawn at $\frac{1}{2}$ and 2. Two points with window SD estimate above 0.5 are not shown.
4.6.4 Relationship between variance of the estimates and mixing

We now examine how the variance of an estimate relates to its mixing behavior, summarized by Brooks’ $D$ statistic. Scatter plots of the SD estimates from the window estimator against the $D$ statistic were drawn for $S_{most}$ (figure 4.18) and $S_{pairs}$ (figure 4.19). For $S_{most}$ the points where the SD is low and $D$ near 0 are due to the low variability of the estimates at some loci, as discussed in section 4.6.2. Increasing SD estimates are seen as $D$ increases from near 0 to 0.35, an effect of the progressive relaxation of the bound on the maximum value of $D$ imposed by the low variability of the estimates. Beyond 0.35 the trend is reversed and the SD estimate decreases as $D$ increases. The lower end of the range of $D$ for $S_{pairs}$ is higher than with $S_{most}$ and the relationship between $D$ and the SD estimate is decreasing over the whole range. The negative relationships reflect an improvement in precision with better mixing. In both cases the correlations are weak. The slope of the relationship between $D$ and the actual SD is likely to be steeper because the SD tends to be underestimated when mixing is poor and $D$ low. The present results on the relationship between mixing speed and variability across different data for the same sampler have to be distinguished from the results of section 3.8 where the SD and mixing are studied for different samplers on the same data. Conditional on the data the relationship is much stronger.

Returning to the loci where cusum plots were drawn in figures 4.12 to 4.15, we find that the estimated SD at D6S422 is 0.145 for $S_{most}$ and 1.25 for $S_{pairs}$. At D2S396 the SD estimates are 0.011 and 0.045 respectively. The estimates at D6S422 are much higher than at D2S396, but given the large peaks in the cusum path at D6S422 revealing sudden shifts in the chain, even those high estimates may be underestimates if the chain has not sampled the stationary distribution to its full extent.

4.6.5 Influence of marker information on variance of estimates and mixing speed

An attempt was made to understand the influence of the marker information on the variance of the estimates and the mixing speed. We examined the relationship between SD estimates and the $D$ mixing statistic on the one side and summary statistics of the information provided by marker data at the locus on the other side. The summary
Figure 4.18: Window estimate of the SD of $\hat{S}_{\text{most}}$ against $D$ statistic. Two points with SD estimate above 0.5 are not shown.

Figure 4.19: Window estimate of the SD of $\hat{S}_{\text{pairs}}$ against $D$ statistic.
statistics considered are the number of individuals with an observed marker phenotype, the heterozygosity of the marker estimated by the proportion of marker phenotypes that are heterozygous and the product of those two variables, the observed number of heterozygous individuals. There is no visible association between the SD estimates and the observed number of heterozygotes on figure 4.20 or 4.21 and the picture was similar for the other variables not shown here. The actual variance of the IBD sharing statistics conditional on the marker data is by definition dependent on the information in the marker data. The Monte Carlo variance being a function of that variance, a positive relationship was expected between the SD estimates and summary measures of marker information. The failure to observe such a relationship may be explained by an underestimation of the variance worsening as marker information decreases, or by the summary variables not capturing the features of the marker data which have a determinant effect on the variance of the IBD sharing statistics.

![Figure 4.20: Window estimate of the SD of $\hat{S}_{most}$ against number of observed heterozygotes.](image)

The $D$ measure of mixing shows no association with the observed number of heterozygotes either (figures 4.22 and figures 4.23). The link between the statistic $D$ and the marker information is more indirect than with the variance, and the lack of association less surprising.
Figure 4.21: Window estimate of the SD of $\hat{S}_{pairs}$ against number of observed heterozygotes.

Figure 4.22: $D$ statistic for $\hat{S}_{most}$ against number of observed heterozygotes.
4.6.6 Convergence at the level of condensed IBD configurations

The frequencies of IBD sharing patterns at individual marker loci over batches of 2000 realizations were also recorded with the aim of tracking statistics closer to the states of the meiosis indicator matrix $S$. Those IBD sharing patterns are condensed versions of the IBD configurations, where only the labels of the three most represented alleles are preserved and the other allele labels replaced by the null value 0, the largest sets of genes IBD being those of greatest interest. When two or more alleles are tied in term of number of copies, the allele with the lowest label is retained. An example of conversion from an IBD configuration to its condensed form is given below. Remembering that the genes in a pair are unordered, relabeling a gene 0 may involve swapping the two genes in a pair to insure that the gene labeled 0 is on the left of a gene with a higher label.

| IBD configuration | 1 2 1 3 3 4 3 5 4 5 |
| Condensed configuration | 0 1 1 3 3 4 0 3 0 4 |

The justification for that data reduction is that IBD configurations between 20 or more individuals observed in a MCMC run almost always differ from all the others for some of the numerous pairs and triples of identical genes that they contain while their largest
subsets are common to multiple configurations.

Table 4.4 shows the counts of condensed IBD configurations at locus D6S422 where cusum plots reveal poor mixing. This locus illustrates the symptoms of poor mixing: the proportions of realizations of condensed IBD configurations vary greatly from one batch of 2000 realizations to the next. At D2S396 where mixing appears good on a cusum plot the distribution is much more stable from batch to batch (table 4.5).

<table>
<thead>
<tr>
<th>Rank</th>
<th>1-2000</th>
<th>2001-4000</th>
<th>4001-6000</th>
<th>6001-8000</th>
<th>8001-10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>283</td>
<td>58</td>
<td>208</td>
<td>331</td>
<td>225</td>
</tr>
<tr>
<td>2</td>
<td>270</td>
<td>65</td>
<td>143</td>
<td>276</td>
<td>243</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>370</td>
<td>114</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>123</td>
<td>11</td>
<td>95</td>
<td>116</td>
<td>101</td>
</tr>
<tr>
<td>5</td>
<td>68</td>
<td>215</td>
<td>110</td>
<td>3</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 4.4: Counts of the 5 most frequent condensed IBD configurations at locus D6S422.

<table>
<thead>
<tr>
<th>Rank</th>
<th>1-2000</th>
<th>2001-4000</th>
<th>4001-6000</th>
<th>6001-8000</th>
<th>8001-10000</th>
<th>10001-12000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>150</td>
<td>128</td>
<td>146</td>
<td>106</td>
<td>119</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>121</td>
<td>116</td>
<td>119</td>
<td>113</td>
<td>125</td>
<td>119</td>
</tr>
<tr>
<td>3</td>
<td>93</td>
<td>107</td>
<td>91</td>
<td>95</td>
<td>82</td>
<td>102</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>65</td>
<td>60</td>
<td>80</td>
<td>87</td>
<td>77</td>
</tr>
<tr>
<td>5</td>
<td>53</td>
<td>67</td>
<td>90</td>
<td>85</td>
<td>72</td>
<td>83</td>
</tr>
</tbody>
</table>

Table 4.5: Counts of the 5 most frequent condensed IBD configurations at locus D2S396.

4.6.7 Variability between runs

Multiple MCMC runs started from distinct initial meiosis indicator configurations may provide a confirmation of a diagnosis of slow mixing within a run or detect unsuspected failure to converge when they produce divergent estimates of expectations or stationary distributions. A second MCMC run was performed on the marker data from the pedigree GTAS2 on a few chromosomes as a validation of the diagnostic results from a single run.

Chromosome 6 was arbitrarily selected for a second run without regard to the estimates in the first run. The initial meiosis indicator matrix for that second run of length
4000 was generated using the procedure of section 3.5 from a different seed locus than in the first run.

The marginal distributions of condensed IBD configurations of the two runs at each locus were used to compare the runs. Configurations with an empirical frequency of less than 0.001 in both runs, or whose empirical frequency was ranked lower than the 100th in both runs were collapsed into one configuration. With two runs, a $\chi^2$ distance was computed between the two distributions. It is defined as

$$\chi^2_{k-1} = (N_1 + N_2) \sum_{j=1}^{k} \frac{(\hat{p}_{1j} - \hat{p}_{2j})^2}{\hat{p}_{1j} + \hat{p}_{2j}}$$

where $N_1$ and $N_2$ are the numbers of realizations sampled from the two runs, $k$ is the number of configurations and $\hat{p}_{1j}$ and $\hat{p}_{2j}$ the estimated probabilities of the $j^{th}$ condensed IBD configuration. One of those is the collapsed configuration; it introduces a slight bias downward in the $\chi^2$ distance. The ratio of the $\chi^2$ distance over its degrees of freedom, $\chi^2_{k-1}/(k-1)$, is reported as a standardized measure.

The estimated values from each of the two runs of the statistic $S_{most}$ for the whole set of 24 affected individuals at the marker loci on chromosome 6 are shown on figure 4.24 with pointwise 95% confidence intervals. At many loci the confidence intervals are too narrow to be visible on the plot. The estimates from the two runs are almost identical at some loci but far apart at others, and in those cases the estimate for one run lies far outside the confidence interval for the other, casting a doubt on the validity of the variance estimates used to construct the intervals. The standardized $\chi^2$ distance is plotted on figure 4.25. The distance between the two runs is highest at the seed loci of the runs, implying that a dependence on the starting point of the chains remains. The proportion of condensed IBD configurations that are seen in both runs among the $k$ IBD configurations appearing in either run is less than 10% at 5 out of 20 loci including two where there is no overlap at all. This implies that the chains are sampling distinct regions of the space of meiosis indicators. The difference between the estimates of $S_{most}$ is not strongly correlated with the distance between the runs. The two figures illustrate that even disjoint sets of IBD configurations can produce very similar estimates of the conditional expectation of an IBD sharing statistic.
An interesting question is whether mixing diagnostics applied to summary statistics for a single run predict divergence between multiple runs. There are not enough loci where the standardized $\chi^2$ distance has been computed and the trace of the observations from one run recorded to measure the correlation. The standardized $\chi^2$ distance at D6S422 is among the highest (285) aside from the seed loci. This is a locus where the cusum plots for $S_{most}$ and $S_{pairs}$ in the first run revealed poor mixing and the frequencies of the condensed IBD configurations sampled most often exhibit large fluctuations between batches in table 4.4. However, other loci have poor apparent mixing and low standardized $\chi^2$ distance or fast apparent mixing and high standardized $\chi^2$ distance.

![Figure 4.24: Estimate of the conditional expectation of the statistic $S_{most}$ on the 24 affected individuals of the pedigree GTAS2 for chromosome 6 in two MCMC runs.](image)

### 4.6.8 Conclusion of the convergence monitoring

The profile of mixing speed drawn by the cusum plots of sequences of realizations of IBD sharing statistics at loci in the genome scan range from the near random path associated with fast mixing to the long drifts interrupted by abrupt turning points revealing poor mixing, often for statistics recorded within the same MCMC run. Pictures of good mixing behavior can however be misleading if the sampler has not moved through all the
Figure 4.25: $\chi^2$ distance between the marginal distributions of condensed IBD configurations at every marker locus on chromosome 6 in two MCMC runs.

high probability regions of the distribution. Comparison of the statistic values and IBD sharing patterns between two runs for one chromosome revealed serious differences, even at loci where a single run gave an appearance of fast mixing.

In the runs that we performed the estimated variance of IBD sharing statistic estimates was small compared to the variance of the null distribution of the statistics in the pedigree. The variance estimates tend to be higher at loci where monitoring of the statistics reflects poor mixing, but those higher estimates may not be valid and in fact underestimate the variance because the realizations from the chain may not form a representative sample from the marginal distribution $P[S_l|\mathbf{Y}]$ at a locus where mixing is poor. Neither is an appearance of fast mixing a guarantee that the realizations are sampled from the entire distribution $P[S_l|\mathbf{Y}]$ and the variance estimates are correct. The variance of statistic estimates and the mixing behavior of a MCMC sampler depend on the constraints imposed on the meiosis indicators by the marker data. Simple summary measures of marker information were however not found to predict mixing behavior or variance of the estimates.
In summary, we find evidence that the hybrid MCMC sampler in most instances does not mix fast enough to draw a sample from the distribution $P[S|Y]$ over the length of the runs that it was feasible to perform. The observed low variance estimates do not guarantee actual convergence of the estimates.

4.7 Inference on IBD sharing at the GLC1A locus

A STOP mutation in the myocilin gene at the GLC1A locus was found in 9 out of 24 affecteds of the pedigree GTAS2 (Craig et al. [7]). Since the mutation is rare it was presumably inherited from a common ancestor. This result gives us an instance where the IBD status of a group of affecteds is known to test the IBD sharing inference from MCMC sampling.

Ignoring the information on the GLC1A locus, we estimated the largest subset of affected individuals sharing a gene IBD at marker loci in the region of the GLC1A locus using the data on the 34 markers phenotyped on chromosome 1. The 9 carriers of a STOP mutation are very likely to form the largest subset of affecteds sharing a gene IBD at the GLC1A locus; the most probable largest subset estimated from MCMC runs should therefore match the set of 9 mutation carriers.

Table 4.6 shows results obtained at four marker loci in the neighborhood of the GLC1A locus from the MCMC run of length 14,000 performed on the marker data on chromosome 1 as part of the analysis of the genome scan. The probability estimates are the proportions of realizations in the MCMC run where the displayed subset of individuals was the largest subset sharing a gene IBD, marginally at each locus.

In that first run, the 9 affected individuals carrying the mutation are predicted to share IBD at the locus D1S218 flanking GLC1A and 8 out of 9 at D1S196, the other flanking locus. The exclusion of individual 32 from the set of individuals sharing IBD at D1S196 could be explained by a recombination between GLC1A and D1S196. Some non-carriers of the mutation are predicted to share IBD with the mutation carriers at one of the loci flanking GLC1A, but none at both. The estimated probabilities of those most probable largest subsets are above 0.4 at all loci except D1S2878. Those high probabilities would
Table 4.6: Highest probability largest subset of affecteds sharing an allele IBD for each of 5 loci around the GLC1A locus conditional on data at 34 marker loci on chromosome 1 in a first MCMC run. Individuals predicted to share IBD: + marker genotype present, o marker genotype missing.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Affecteds carrier of the mutation</th>
<th>Affecteds non-carrier</th>
<th>Estimated prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1S2878</td>
<td>+ + + + + + + + + +</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>D1S196</td>
<td>+ + + + + + + + + +</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>GLC1A</td>
<td>D1S218</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ + + + + + o + + o</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D1S238</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ + + + + + + + + +</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D1S413</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ + + + + + + + + +</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A second run of 6000 observations was performed on the same data starting from a different initial configuration. Results are presented in table 4.7. The largest subsets of affecteds inferred in that second run differ from those in the first run. At the marker loci surrounding GLC1A, 2 of the 9 mutation carriers are excluded from the largest subset of individuals sharing IBD. Fewer affected non-carriers are included into the largest subsets at most loci. The largest subsets in the second run are sampled with very high frequency at the three middle loci, but those frequencies are inconsistent with the results from the first run where the same subsets did not rank as the most probable largest subsets. Observed frequencies in both runs are therefore bad estimates of the actual probabilities of the subsets. These results are another illustration that on the pedigree GTAS2 the MCMC sampler may sample distinct regions of the space of meiosis indicators corresponding to different IBD configurations in runs started from different initial states. The first run gave results in good agreement with the known IBD sharing between 9 affected individuals at the GLC1A locus in the middle of the region. The results from the second run do not agree so well with the known IBD sharing, but the estimated largest subset of affecteds sharing IBD comprises most of the individuals sharing IBD at GLC1A. In that sense the inferred subsets in the
second run are close to those from the first run, and conclusions from the second run are not totally misleading.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Affecteds carrier of the mutation</th>
<th>Affecteds non-carrier</th>
<th>Estimated prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15 17 18 19 20 32 46 96 113</td>
<td>34 56 57 88 654 10 others</td>
<td></td>
</tr>
<tr>
<td>D1S2878</td>
<td>+ + + + +</td>
<td>+ +</td>
<td>0.22</td>
</tr>
<tr>
<td>D1S196</td>
<td>+ + + + +</td>
<td>+ +</td>
<td>0.77</td>
</tr>
<tr>
<td>GLC1A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1S218</td>
<td>+ + + + +</td>
<td>+ +</td>
<td>0.69</td>
</tr>
<tr>
<td>D1S238</td>
<td>+ + + + +</td>
<td>+ +</td>
<td>1.00</td>
</tr>
<tr>
<td>D1S413</td>
<td>+ + + + + +</td>
<td>+ +</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 4.7: Highest probability largest subset of affected sharing an allele IBD for each of 5 loci around the GLC1A locus conditional on data at 34 marker loci on chromosome 1 in a second MCMC run. Individuals predicted to share IBD: + marker genotype present, \(\circ\) marker genotype missing.

4.8 Results of the IBD sharing analysis of the genome scan

In this section we present the estimates of conditional expectations of IBD sharing statistics computed from the marker data observed in a genome scan on the pedigree GTAS2. The estimates were obtained from a single MCMC run on each chromosome. Given the assessment that the chains often do not converge over the lengths of runs that it was feasible to perform, it is likely that some of the estimates are far from the true value. Those results contain nevertheless valuable information on the behavior of the statistics and the MCMC methods applied to compute them.

Estimates of the conditional expectation of the statistics \(S_{\text{most}}\) and \(S_{\text{pairs}}\) at the 401 marker loci included in the genome scan for the full set of 24 affected individuals and the reduced set of 15 affecteds who do not carry a STOP mutation in the myocilin gene are plotted on figures 4.26 to 4.29.

The statistic generalizing \(S_{\text{most}}\) to allow for heterogeneity between the two main
sub-pedigrees formed by the descendents of 2007 and 2011 was also considered for the full set of 24 affecteds. With incomplete marker data, IBD configurations where the largest subset of IBD genes spans the two sub-pedigrees and IBD configurations where it is restricted to the largest sub-pedigree may both be compatible with the marker data. This indeed occurs in the analysis of the genome scan on GTAS2 at 214 loci. At the remaining 187 loci all the realizations sampled had a largest subset restricted to the pedigree below 2011. One analysis approach at loci where both cases appear would be to compute the form of the statistic applicable to each sampled realization as described in section 4.3.1, and weigh the evidence from the two forms of the statistic by the estimated probability of each. However, since meiosis vectors giving a largest subset limited to the sub-pedigree below 2011 were sampled more frequently than meiosis vectors giving a largest subset overlapping the two sub-pedigrees at all but 8 loci, we opted to report only the statistic computed from the realizations where the largest subset of genes IBD is restricted to the sub-pedigree below 2011. The statistic computed in that case was the sum of $S_{\text{most}}$ for the two sub-pedigrees. The estimates of that statistic are plotted on figure 4.30.

When all 24 affecteds are considered, $\hat{S}_{\text{most}}$, $\hat{S}_{\text{pairs}}$ and the estimate of the sum of $\hat{S}_{\text{most}}$ in the two sub-pedigrees reach their highest values of 13.6, 97.7 and 16.0 respectively at the same locus on chromosome 3. On the subset of 15 affecteds without the STOP mutation in the myocilin gene, $\hat{S}_{\text{most}}$ reaches its maximum value of 7.1 and $\hat{S}_{\text{pairs}}$ reaches its maximum value of 28.2 at two adjacent loci roughly 20cM from the peak for the full set of 24 affecteds. Those highest estimates remain below the critical values for a genomewide scan. Critical values were however set assuming that IBD sharing is observed, and may be too high for the conditional expectations of the statistics, as discussed in section 2.6.

4.8.1 Comparison of the distribution of IBD sharing statistics estimates at marker loci of the genome scan to the null distribution of the statistics

As emphasized in section 2.6 the null distribution of $Z_l$ depends on the distribution of genotypes of markers included in the computation of $P[S_l | Y]$, and values observed at different locations unlinked to any disease gene would not be identically distributed. It is still
useful to consider the empirical distribution of estimates $\hat{Z}$ at the marker loci of a genome scan as estimating a mixture distribution of $Z_l$ representing a background distribution of $Z$ given the marker data on the genome.

Figures 4.31 and 4.32 show quantile-quantile plots of the empirical distribution of $\hat{Z}$ at the genome scan loci against the distribution of $Z$ for $S_{\text{most}}$ and $S_{\text{pairs}}$ computed on the full set of 24 affected individuals. The points fall close to a straight line indicating that the mixture distribution of $Z_l$ has a shape close to the distribution of $Z$. The ratio of the SD of the two distributions is 0.95 for $S_{\text{most}}$ and 0.93 for $S_{\text{pairs}}$, meaning that the dispersion of $Z_l$ is less than the dispersion of $Z$ assuming $\hat{Z}$ is a good estimate of $Z_l$. The SD ratios are the slopes of the lines on the quantile-quantile plots.

The only two points standing above the line on figures 4.31 and 4.32 are the two
close neighbors with the highest values of the IBD sharing statistics identified at section 4.8. Assuming that the distribution of $\tilde{Z}_l$ at those loci is close to the mixture distribution for all loci this would be indication of unusually high sharing. Since that assumption cannot be verified, it is not possible to formally evaluate the significance of the results. The points visually detected on the quantile-quantile plots and corresponding to the highest values of the statistic computed on 24 affecteds are closer to the critical value than the maximum for 15 affecteds. The region around the peak for the 24 affecteds was therefore selected for further investigation.

4.8.2 Follow-up of the region with the highest signal

A second MCMC run of 4000 realizations was performed on the data from chromosome 3 using the locus with the highest values of the statistics as the seed locus to
generate the initial configuration. Contrary to the first run, the estimates from that second run were not higher than the expectation of the statistics under the null. The estimates of the conditional expectation of $S_{pairs}$ for four loci in the region labeled $M_1$ to $M_4$ for the two runs are shown in figure 4.33.

Loci $M_2$ and $M_3$ are close neighbors. The presence of two close consecutive markers reduces the probability of meiosis switches at one of the two loci by the locus sampler, and this may negatively impact the mixing speed. Other runs of length 4000 were performed where either $M_2$ or $M_3$ was removed. The estimates $\hat{S}_{most}$ for these runs are also plotted on figure 4.33. They are close to the estimates from the second run with all the markers, showing no elevated sharing.

The first run on chromosome 3 was extended from 10,000 to a total of 20,000
realizations in the hope that the chain would move to the regions of the space of meiosis indicators it had not explored yet. Estimates after 20,000 realizations remained close to what they were after the first 10,000 realizations. Sampling 20,000 realizations required about 155 hours of computing time, almost a full week. A cusum plot of the sampled values of $S_{most}$ at locus $M_4$ (figure 4.34) shows signs of stickiness, but the global extremum, about -400, is only a few times larger than the extremum of the shuffled sequence. Mixing failure becomes apparent only by comparing the two runs. The values sampled in the second run do not even overlap with those in the first run, as we see on a sequential plot of $S_{most}$ for the first 4000 realizations of the first run and the entire second run (figure 4.35).

The inability of the sampler to cover the whole conditional probability distribution $P[S|Y]$ in a reasonable amount of computing time leaves us with two irreconcilable results. In an effort to resolve which of the two runs sampled the highest mode of the distribution,
Figure 4.30: Estimates of the conditional expectation of the sum of $S_{most}$ in two sub-pedigrees of GTAS2 with 19 and 5 affected individuals. The dashed line represents the critical value for a genomewide scan computed with the method of section 2.3.1.

The sampled states were recorded in two runs of length 4000 where the meiosis indicator matrix was limited to the four loci $M_1$ to $M_4$ due to the disk space required to store the matrices. One run was started from the initial configuration of the first run (leading to high estimates) and the other from the initial configuration of the second run (leading to estimates close to the null expectation). The joint probability $P[S, Y]$ was computed for each of the sampled $S$. Box plots of the distribution of $\log(P[S, Y])$ in the two samples are presented in figure 4.36. The states from the run with high values of the IBD sharing statistics tend to have higher probability than those from the run with statistics close to their expectation. There is however considerable overlap between the distributions of the probabilities, preventing us from concluding that one of those two runs in confined to a low probability region of the space. The relative importance of the probability mass under the two modes being unknown, it is not possible to determine the relative weight to give to the estimates from each of the runs.
Figure 4.31: empirical distribution of $\hat{S}_{most}$ at 401 marker loci against null distribution of $S_{most}$ for 24 affecteds in the pedigree GTAS2.

Figure 4.32: empirical distribution of $\hat{S}_{pairs}$ at 401 marker loci against null distribution of $S_{pairs}$ for 24 affecteds in the pedigree GTAS2.
Figure 4.33: Estimates of the conditional expectation of $S_{most}$ for 24 affecteds at four markers on chromosome 3 from several MCMC runs. The dotted line is the expected value under the null.

Ultimately, evidence in favor of or against the inference that 13 or more affecteds share genes IBD will be provided by phenotypes at additional markers in the region. With a denser map of markers fewer recombination events occur between adjacent markers and it becomes easier to infer sharing of segments of chromosome IBD. The follow-up with additional markers has been undertaken by investigators of the GIST. At the time of this writing, the data collection is in progress and the data are not yet available.
Figure 4.34: Cusum plot of the first 18,000 consecutive realizations of $S_{\text{most}}$ on 24 affecteds at locus $M_4$.

Figure 4.35: Sampled values of $S_{\text{most}}$ for 24 affecteds at $M_4$ from two MCMC runs.
Figure 4.36: Distribution of $\log(P[S|Y])$ for the sampled realizations of $S$ for four marker loci on chromosome 3 in two MCMC runs, one started from an initial state leading to high values of IBD sharing statistics (high sharing) and the other from an initial state leading to values of IBD sharing statistics near their expectation (low sharing).
Chapter 5

Discussion and conclusion

In large multigeneration pedigrees the information on the transmission of genes from the founders down to the youngest descendants provided by genetic markers is incomplete, mostly because the marker phenotypes of ancestors are unobserved. The pedigree GTAS2 shown in figure 4.1 provides a good example. The resulting uncertainty on the ancestral origin of alleles and the IBD configuration of sets of genes can seriously compromise the power of linkage analysis using a single marker at a time.

With incomplete information, a conditional probability given phenotype data at multiple marker loci can be attached to the gene transmission patterns encoded in meiosis indicators at a particular point of a genetic map under the framework of a hidden Markov model combining a model of the crossover process during meiosis and a model of the marker genotype distribution in the population, both detailed in section 2.5. Under this HMM, the complexity of multilocus computations is linear in the number of loci. Modeling assumptions are also needed because the amount of data available to estimate the high dimensional marker genotype distribution in the population is usually fairly limited. The probabilities inferred under the HMM are useful approximations in the analysis of small pedigrees where combining the information from multiple markers increases the power of linkage statistics. We hypothesized that multilocus analysis would provide a gain in information and power over a single marker analysis in large pedigrees as well.

The challenge in applying multilocus analysis to large pedigrees is computational, since the complexity of the task grows exponentially with the number of meioses and hence
with the size of the pedigree. Given the impracticality of exact algorithms, we turned to Markov chain Monte Carlo to perform approximate computations. Building on the extensive body of work on the application of MCMC to pedigree analysis, we implemented two previously described meiosis indicator samplers, the whole meiosis and whole locus Gibbs samplers described in sections 3.1 and 3.4, and developed two Metropolis samplers that we named the parental sampler and the chain restarting sampler, described in sections 3.3 and 3.6. By flipping all the indicators from a parent, the parental sampler allows to step between meiosis indicator states that would not communicate by updating a single meiosis at a time due to constraints in the phenotype data in nuclear families investigated in section 3.2. States may however fail to communicate under the parental sampler due to constraints involving data from multiple nuclear families in a pedigree. The chain restarting sampler seemed particularly promising in that it updates all meiosis indicators at once and shares with the locus sampler the property of generating an irreducible chain. It turned out however to have an acceptance probability too low to be applied to very large problems where it would be most useful.

An hybrid sampler can be constructed by alternating steps of different MCMC samplers. Little is known about the combinations of samplers and the proportions of steps by each sampler that optimize precision and accuracy. We investigated hybrid samplers empirically in section 3.8 by comparing their bias and variance after a run of a set length on two test problems for which the computation of statistics of interest is pushing exact computing algorithms to their limits. The main conclusion from that study is that combining more types of samplers tends to decrease variance and bias. The chain restarting sampler reduces the variance of the estimates for a fixed run length on a small pedigree but on a larger one the acceptance probability is so low that it has the opposite effect. The locus and chain restarting samplers are the most time consuming, especially on the larger test pedigree, but because other samplers are not irreducible, one of those samplers must be included to insure that the hybrid sampler is theoretically irreducible.

The information on gene transmission at a locus extracted from marker data is used to test for linkage to a disease locus. The IBD sharing approach is based on the observation that affected relatives have a higher probability of sharing genes IBD at or near a locus influencing susceptibility to a disease than they have at an unlinked locus under a wide
range of models of the relationship between genotype and disease phenotype (section 2.2). Statistics quantifying an aspect of IBD sharing provide a means to test the null hypothesis without specifying a genetic model for the trait that would likely be misspecified, especially when multiple genes contribute to a disease. The selection of an IBD sharing statistic is however ad hoc. In section 4.3 we designed IBD sharing statistics for the linkage analysis of glaucoma in the large pedigree GTAS2 based on hypotheses on the genes involved that were guided by the structure of the pedigree. Such an approach, which was intended to produce statistics powerful to detect glaucoma genes in GTAS2, is in some sense related to specifying a genetic model for the trait. We established that the statistic $S_{\text{most}}$ is equivalent to the likelihood ratio to test the null hypothesis that the locus where the IBD configuration is observed is unlinked to the disease locus against the alternative that it coincides with the disease locus under a quasi-dominant genetic model for the trait when the disease allele frequency goes to 0 and the risk ratio for a carrier over a non-carrier of the disease allele is large.

The results of a simulation study on the pedigree GTAS2 in section 4.4.1 show that the statistic $S_{\text{most}}$ and its extension allowing sharing of distinct genes in two sub-pedigrees had a power roughly equal to the generic statistic $S_{\text{pairs}}$ under scenarios of disease etiology compatible with the formulated hypotheses, and assuming complete inheritance information. Only under allelic heterogeneity between the two sub-pedigrees did the extension of $S_{\text{most}}$ designed for that type of heterogeneity performed better than $S_{\text{pairs}}$. We must conclude that it is preferable to use $S_{\text{pairs}}$ given the results from studies demonstrating it performs well under a range of models, especially recessive models where $S_{\text{most}}$ would have almost no power.

The power of both $S_{\text{most}}$ and $S_{\text{pairs}}$ is about 80% in three of the four scenarios considered, and much less in the last one where the gene at the locus tested is often responsible for few of the disease cases. The actual level of power with incomplete marker data is lower, and it depends critically on our ability to extract the information on gene transmission from the data. The simulation study of section 4.4.2 reveals that the conditional expectation $\tilde{Z}$ of an IBD sharing statistic $Z$ near a disease locus increases when the number of linked markers used goes from one to three and from three to six, confirming the presumption that more information on gene transmission can be extracted by using more marker loci. The difference between the conditional expectation computed using six
marker loci and the actual value of the statistics remains important, showing the limits of
gene transmission inference in large pedigrees with missing data on the ancestors when the
marker density is around one marker every 10 cM, a typical figure for a genome scan. The
impact on power can be mitigated if the null distribution of $\tilde{Z}$ is used to assess significance
instead of the distribution of $Z$ because of its lower dispersion, but that alternative is not
computationally practical.

When $\tilde{Z}$ is computed at all the loci of a genome scan, most loci are unlinked to
any disease locus and the values of $\tilde{Z}$ at those loci reflect the background distribution of $\tilde{Z}$
given the data observed over the whole genome. A quantile-quantile plot of the empirical
distribution of $\tilde{Z}$ at the genome scan loci against the null distribution of $Z$ like the one
presented in section 4.8.1 permits to judge the agreement between the shapes of the two
distributions, and evaluate their scale ratio. Some points may stand above the line on which
most of the points fall without having a value of $\tilde{Z}$ reaching the genomewide critical value
computed from the distribution of $Z$ if the scale of the null distribution of $\tilde{Z}$ is lower than
that of the distribution of $Z$. Those points may represent a true signal, provided that the
null distribution of $\tilde{Z}$ at the corresponding loci is close to the background distribution, an
assumption difficult to check and that won’t hold if for instance the quality of the marker
data varies between regions.

An important limitation of MCMC methods is the absence of guarantee that ob-
servations are actually drawn from the target distribution and that estimates have truly
converged after a finite run length. Monitoring the sequence of realizations using tools like
the cusum plot of section 3.7.1 may provide clues that a sampler is mixing too slowly for
the estimates to have converged. Serious shortcomings were detected when assessing the
convergence of the hybrid sampler applied to the data for whole chromosomes on the large
pedigree GTAS2 (section 4.6.2). After runs taking up to one week of computing time the
estimated SD of the estimates of univariate statistics were small compared to the SD of the
null distribution of the statistics, but signs of slow mixing in the sequence of observations
at a large fraction of the loci hinted that the SD was underestimated and that the estimates
could be far off the quantity estimated. Second runs started from different initial states
confirmed the failure of the estimates to converge in many suspected cases and also gave
divergent estimates at loci where bias remained unsuspected based on the output from a
single run, for instance in a region where IBD sharing statistics peaked in the first run studied in section 4.8.2. Inspection of more detailed patterns of IBD sharing, for example the largest set of individuals sharing a gene IBD in the region of chromosome 1 near the GLC1A locus (section 4.7), revealed almost no overlap between the patterns observed in two runs. In that exercise intended to verify whether a subset of affecteds carrying a mutation presumably IBD could be inferred using the incomplete data at nearby markers, the subsets observed in two runs were distinct but nonetheless similar. That similarity gives hope that the estimates from MCMC runs capture the important information even when different regions of the state space are sampled. However, in another example in a region of chromosome 3 (section 4.8.2) the estimates from two runs were greatly different.

The failure to reach convergence in the analysis of the data on the pedigree GTAS2 prompted the development of new MCMC samplers. The chain restarting sampler and the parental sampler showed some promises but our other attempts not reported in this thesis failed to improve performances even in small pedigrees. Without MCMC samplers performing well in very large multilocus problems, the problems on which results from the current MCMC samplers are reliable appear to be limited to pedigrees with about 50 non-founders, larger than the pedigree size on which exact multilocus analysis can be performed using the Lander-Green algorithm but smaller than many multigeneration pedigrees collected for genetic linkage studies.

The estimates of conditional expectations of IBD sharing statistics on the pedigree GTAS2 reported in section 4.8 never exceeded the critical values for a genomewide scan, and it is natural to ask whether this is the result of the lack of power of the analysis, the absence of a gene conferring strong predisposition to POAG, or an artifact of the MCMC sampling. In large pedigrees, genome regions where an important proportion of individuals have alleles of the same type, either because many alleles of this type have been introduced in the pedigree or because an ancestral allele is shared IBD, are most prone to estimation bias with a poorly mixing MCMC sampler due to the large number of compatible IBD configurations. For that reason, it may be that some regions with high IBD sharing were missed. However, the high proportion of missing data is likely to be the most important factor to explain the absence of significant signal. The results of section 4.4 illustrate well that $\bar{Z}$ is much lower than $Z$ at a locus where sharing is high. It is also possible that mul-
tiple genes are responsible for the cases of POAG in the pedigree. In that case the power to detect any one of them with only 24 affecteds would be very low.

In light of the work in this thesis, we are pessimistic about the prospects of developing MCMC samplers producing reliable inferences on large pedigrees. The currently available methods may detect regions where the marker data are such that the presence of a disease gene is possible. Then, any ambiguity on IBD sharing must be resolved by collecting data at more markers or observing the phenotypes of more individuals when possible. In cases where IBD sharing can be inferred with confidence, exploiting the information on the sharing of alleles jointly by multiple affected individuals holds promises, but care must be taken to allow for genetic heterogeneity between parts of a large pedigree when studying common traits.
Bibliography


Appendix A

Computing the probability of marker data given the meiosis indicators at a locus

We present here the efficient algorithm introduced by Sobel and Lange [52] and Kruglyak et al. [32] to compute $P[Y_i|S_l]$ at a locus $l$. To simplify the notation, we drop the subscript $l$.

Let $a = (a_1, \ldots, a_{2f})$ be a vector of allele types assigned to the $2f$ genes of the founders of the pedigree. The idea is to represent by a graph the restrictions imposed by the observed marker phenotypes on the vectors $a$ that can be assigned to the founder genes. The algorithm extracts from the graph only the vectors $a$ compatible with the marker data, and $P[Y|S]$ is obtained by summing over the compatible vectors.

IBD copies of a pedigree founder gene find themselves paired with copies of other founder genes in the descendents to whom the founder genes have been transmitted following the segregation events encoded in the meiosis indicators $S$ at the locus. A founder graph is defined as a graph whose vertices are the founder genes and whose edges connect the genes appearing together in an individual whose marker phenotype, i.e. unordered genotype, is observed. The edges are labeled with the unordered genotype of the corresponding individuals.
As an illustration of the concepts of the algorithm, we apply it to compute the probability of an instance of gene transmission in the pedigree depicted on figure A.1. The gene transmission is represented by what Sobel and Lange [52] call a descent graph, shown in figure A.2(a). Following the convention that paternally inherited genes are on the left, and placing the non-founders in increasing order, this realization of inheritance is encoded into the meiosis vector \( S = (0, 0; 1, 1; 0, 0; 0, 0; 0, 0; 1, 1) \).

![Figure A.1: Pedigree and marker data used in the example of computation of the probability of marker data given the meiosis indicators at a locus.](image)

The founder graph associated with this realization of inheritance is represented in figure A.2(b). An edge connects genes 3 and 5 because they both pass through individual 12 whose marker phenotype is observed. Then, gene 3 is transmitted to individuals 21 and 22 along with gene 1 from individual 11, so an edge is traced between genes 1 and 3, and so on.

A founder graph splits into \( m \) connected components labeled \( C_1, \ldots, C_m \). By construction, the founder genes assigned to different components are found in the genotype
Figure A.2: Examples of descent graph and its corresponding founder graph. Paired genes form the genotype of a pedigree member. Founder genes are numbered from 1 to 8.
of different observed individuals. Under assumption 4 of Hardy-Weinberg equilibrium, the allele types of different founder genes are independent, allowing to compute the probability of each component separately and multiply the component probabilities.

It may happen that some genes never pass through observed individuals. Their vertices form singleton components, and any allele type can be assigned to them. In the founder graph of figure A.2(b), the vertex 2 forms a singleton.

After constructing the graph, the following algorithm finds the allele assignments compatible with the observed marker phenotypes for each non-singleton component:

1. Pick an arbitrary vertex in the component and determine its set of compatible allele types, the intersection of the genotypes attached to the edges incident to the vertex. This is saying that a gene can only be of an allele type that is observed in all the individuals where the gene is present. The set of compatible allele types contains at most two elements, since the gene is found in at least one genotype and no more than two allele types are present in a genotype. If the intersection of the genotypes is empty, there is no compatible assignment for the component and hence for the graph, then abort the algorithm.

2. If the set of compatible allele types of the initial vertex contains one element, assign that allele type to the vertex and perform step 3 once. If the set contains 2 elements, assign one of the allele types to the vertex, perform step 3, store the allele assignment it returns, assign the other allele type to the vertex and repeat step 3.

3. Selecting the allele type of one vertex forces the adjacent vertices to be of the other allele type in the genotype attached to the edge joining them to the original vertex. Visit each vertex in turn and check that the allele to be assigned to the vertex is in the intersection of its genotypes. If, at any point, an incompatible allele type is encountered, no compatible assignment can be obtained from the allele type we started from. Abort and return an empty list. Else record the allele type assigned to the vertex. If all vertices have been assigned a compatible allele type, return the list of recorded allele types.

At the end of the execution, we get a set of compatible allele assignments for each
connected component and we denote the sets by $A_1, \cdots, A_m$. Table A.1 gives the sets of compatible assignments for the components in the founder graph of figure A.2(b).

<table>
<thead>
<tr>
<th>Graph component</th>
<th>Allele assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2)</td>
<td>(a), (b), (c), (d)</td>
</tr>
<tr>
<td>(1,3,5)</td>
<td>(a,b,a), (b,a,b)</td>
</tr>
<tr>
<td>(4,6,7,8)</td>
<td>(a,b,c,d)</td>
</tr>
</tbody>
</table>

Table A.1: Compatible allele assignments for the components of the example founder graph.

Each set $A_i$ contains 0, 1 or 2 assignments, except for singleton components. An empty set for at least one of the components means that for the meiosis vector $S$ we are considering, there is no compatible allele assignment and therefore $P[Y|S] = 0$. The only global compatible assignments are those in the Cartesian product of the sets $A_1, \cdots, A_m$.

The final step is the computation of the desired probability. We ignore singleton components since their probability is 1. For the other components, let $a_{hi}$ be an element of $A_i$ (a vector of alleles assigned to the vertices of the component $C_i$). Then,

$$
P[a_{hi}] = \prod_{j \in C_i} P[a_j]$$

$$
P[C_i] = \sum_{h:a_{hi} \in A_i} P[a_{hi}]$$

$$
P[Y|S] = \prod_{i=1}^{m} P[C_i]$$

The summation over the $a_{hi}$ contains a maximum of 2 terms. We take the product over $2f$ vertices. The maximum number of operations is therefore $4f$, and we see that the size of the computation scales linearly with the number of founders in the pedigree.
Appendix B

MCMC sampling software

The MCMC samplers described in this chapter have been implemented in a computer program written in ANSI C language called PLASMIN, which stands for Pedigree Linkage Analysis by Sampling of Meiosis INdicators. The program reads three input files: a marker file containing information such as locus order and allele frequency and genetic distance estimates in the format required by the LINKAGE software, a pedigree file listing individuals, their parents, sex, affection status and marker phenotypes in pre-Makeped format and a file specifying the values of the parameters of the MCMC runs such as the proportion of steps of each sampler, the run length and the statistics to compute. The program outputs the mean of IBD sharing and likelihood ratio statistics for the entire sample and for batches of realizations as well as the frequency of IBD configurations observed in the sample. The sequence of observations of linkage statistics can be saved to a text file to be loaded into a general purpose statistical package like S-PLUS for plotting and computation of convergence diagnostics. PLASMIN is available on the World Wide Web at www.stat.berkeley.edu/users/terry/linkage/software.html.