INFORMATION TO USERS

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

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

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

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

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

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Monte Carlo Likelihood Calculation for Identity by Descent Data

Sharon Browning

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

Doctor of Philosophy

University of Washington

1999

Program Authorized to Offer Degree: Statistics

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

Signature Sharon Browning

Date 7/27/99

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Sharon Browning

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

Elizabeth Thompson

Elizabeth Thompson

Reading Committee:

Elizabeth Thompson

Elizabeth Thompson

Krzysztof Burdzy

Philip Green

Philip Green

Date: July 27, 1999

University of Washington

Abstract

Monte Carlo Likelihood Calculation for Identity by Descent Data

by Sharon Browning

Chair of Supervisory Committee

Professor Elizabeth Thompson
Statistics

Two individuals are identical by descent at a genetic locus if they share the same gene copy at that locus due to inheritance from a recent common ancestor. Identity by descent can be thought of as a continuous process along the genome that is the outcome of a highly structured hidden process. The complexity of the structure rules out direct analytic methods for calculating likelihoods in most situations, so that a Monte Carlo approach is required. This thesis presents an approach that applies to many models for the underlying genetic process of crossing-over at meiosis. The method is applied to simulated data in order to examine the amount of information contained in identity by descent data about the true model for the crossing-over process and about the true relationship between the two individuals from whom the data derive. Much of the work is done with idealized continuous identity by descent data, but an extension to the Monte Carlo method is developed that allows analysis of real data. Real data consist of identity (not necessarily by descent) of gene copies at discrete locations along the genome. The method is applied to relationship inference analysis of a real data set.

TABLE OF CONTENTS

List of Figures v

List of Tables vii

Chapter 1: Introduction 1

1.1 Identity by descent data 1

1.2 The process of crossing-over 3

1.3 Relationship inference 5

1.4 Notation 6

Chapter 2: Monte Carlo calculation of likelihoods 8

2.1 The combined crossover process 8

2.2 Monte Carlo algorithm 11

2.2.1 Choice of P^* 12

2.2.2 Modifications to the Monte Carlo algorithm 16

2.3 Relationships with same expected IBD proportion 17

2.3.1 Details of the simulation study 20

2.4 Other approaches to likelihood calculation 21

2.4.1 An exact method for discrete data with Haldane's model . . . 21

2.4.2 A numerical method for half-sib type relationships 22

Chapter 3: Models for crossing-over 24

3.1 Haldane's model 24

3.2 Chiasma process models 27
3.3 Chi-square model 30
3.3.1 Monte Carlo calculation of likelihoods for the chi-square model 31
3.4 Sturt and truncated Poisson models 35
3.4.1 Calculation details for the truncated Poisson model 36
3.5 Kosambi renewal model 38
3.6 Model comparison 39
3.6.1 Details of the simulation study 42
3.7 Sex-specific rates of crossing-over 43

Chapter 4: Discrete data likelihoods **45**

4.1 Discrete data 45
4.2 Monte Carlo 46
4.2.1 Choice of P_A 47
4.2.2 The normalizing constant 48
4.2.3 Sampling realizations of $\tilde{I}$ 49
4.2.4 Accounting for data errors 58
4.2.5 Use of common IBD process realizations 58
4.3 Bilateral relationships 59
4.4 Evaluation of the MCMC procedure 62
4.4.1 Evaluation of the use of common IBD process realizations . . 66

Chapter 5: Data analysis **67**

5.1 Preliminary analysis 69
5.2 Further analysis of twenty pairs of individuals 70
5.2.1 Issues pertaining to relationship inference 71
5.3 Assessing significance and goodness-of-fit 75

5.3.1 Significance with missing data 76
5.3.2 Goodness-of-fit and power 79
5.3.3 Distinguishing between relationships 80
5.4 Beyond Haldane's model 81
5.4.1 Choice of parameters for calculations 83
5.4.2 Results 85
5.4.3 Discussion of results 87

Chapter 6: Conclusion **88**
6.1 Discussion 88
6.1.1 Main achievements of this thesis 88
6.1.2 Summary of results 89
6.2 Future work 91
6.2.1 Map uncertainty 91
6.2.2 Multiple individuals 92

Bibliography **94**

Appendix A: Details of numerical method for half-sib type pedigrees **98**

Appendix B: Probabilities for Sturt and truncated Poisson models **100**
B.1 Chiasma process probabilities for count-location models 100
B.1.1 The truncated Poisson process is a renewal process 101
B.1.2 The Sturt process is not a renewal process 102
B.1.3 The truncated Poisson model is unique in being a count location model and renewal model 103
B.2 Crossover process probabilities 104

Appendix C: Sex-specific map method for Haldane's model **105**

C.1 The Poisson process model for sex-specific rates 105

C.2 The pedigree-average map 106

C.3 Monte Carlo simulation 108

C.4 Monte Carlo likelihoods 109

Appendix D: Approximations to the likelihood based on local properties **113**

D.1 A Markov model 114

D.1.1 Markov approximation to the log likelihood when the coefficients of kinship are equal 116

D.2 Renewal approximation 120

D.2.1 Modeling region length distributions 120

D.2.2 Empirical length distributions 122

D.3 Conclusion 123

LIST OF FIGURES

1.1 Copies of a single chromosome in individuals on a cousins pedigree. . 2
1.2 Cousins' IBD data. 2
1.3 Physical versus genetic distance. 4
1.4 Locations of switches in IBD status. 7

2.1 Crossover processes on half-aunt-niece pedigree. 10
2.2 Combined crossover process for half-aunt-niece. 10
2.3 Sampling under P^* for the half-aunt-niece relationship. 13
2.4 Three relationships with the same expected proportion of IBD genome. 18

3.1 The process of chiasma formation. 28
3.2 Sampling of chiasma process under P^*. 29
3.3 Comparison of renewal densities for the Kosambi and chi-square (M=2) models. 32
3.4 Cumulative distribution of distances between adjacent crossovers under chi-square models. 33
3.5 The combined conversion process. 34

4.1 Two possible IBD processes underlying an example of discrete IBS data. 46
4.2 Definition of n_e. 50
4.3 Definition of m^*. 51
4.4 Definition of d^*. 52
4.5 Definition of d^{**}. 53

4.6 Definition of n^*. 53
4.7 Move proposal and reverse move. 55
4.8 Proposal to add a mid region, and reverse move. 57
4.9 Examples of bilateral relationships. 60
4.10 An inbred relationship. 60
4.11 Estimated time factor curve. 66

5.1 Mennonite families. 68
5.2 Indistinguishable relationships. 74

6.1 Data on multiple individuals. 92

A.1 Examples of half-sib type pedigrees. 99

D.1 Pedigrees used in evaluating the approximations. 114
D.2 Regression approximation. 117
D.3 Markov approximation. 118
D.4 The Markov approximation to log likelihood ratio. 119
D.5 Non-IBD region lengths. 121
D.6 Modeled renewal approximation to the log likelihood ratio. 122
D.7 Empirical renewal approximation to the log likelihood ratio. 123

LIST OF TABLES

3.1 Number of chromosomes needed to distinguish between pairs of models. 40

4.1 Performance measures for the MCMC procedure. 64

5.1 Analysis of twenty pairs of individuals. 72

5.2 Abbreviations for relationship names. 73

5.3 Sample quantiles of LOD(R vs. unrelated|Haldane's model) 77

5.4 Investigation of choice of parameters for crossover model analysis for aunt. 84

5.5 Investigation of choice of parameters for crossover model analysis for half-cousins once removed. 85

5.6 Results for pair 40-1 110-2. 85

5.7 Results for pair 40-1 51-2. 86

5.8 Results for pair 97-1 109-4. 87

ACKNOWLEDGMENTS

I am very grateful to Elizabeth Thompson for all the guidance and encouragement that she has given me during my time at the University of Washington. I would like to thank my reading committee Chris Burdzy and Phil Green for reading my dissertation and providing useful comments, and the remaining members of my committee, Julian Besag, Peter Guttorp and Suresh Moolgavkar, for their helpful comments at my general and final exams. Many thanks also to Aravinda Chakravarti for providing me with the Mennonite data.

This work was supported in part by NSF grant BIR-9305835 and by the Burroughs Wellcome Fund (BWF) through a Program in Mathematics and Molecular Biology (PMMB) training fellowship.

Chapter 1

INTRODUCTION

In this chapter we will describe the nature of identity by descent data. We will discuss the difference between idealized continuous data and real discrete data. The idea of information, both about the relationship between two individuals and about underlying genetic processes, contained in identity by descent data will be introduced.

1.1 Identity by descent data

The relationship between two related individuals is defined by a pedigree (for example the cousins pedigree in figure 1.1) that details the paths of meioses by which DNA from one or more common ancestors could be passed down to either individual. If the DNA of two related individuals could be matched up and compared, we would typically see some regions of gene identity by descent (IBD) — segments of DNA that are identical for the two individuals because both inherited the same sequence from a recent common ancestor (an ancestor in the pedigree defining the relationship between the individuals). For example, (see figures 1.1 and 1.2) cousins have two grandparents in common, and may receive sections of identical DNA from either one.

Identity by descent can be considered to be a zero-one process along each chromosome. At each locus on a chromosome, two related individuals are either IBD or non-IBD.

In chapters 2 and 3 we will consider an idealized view of identity by descent data in which IBD status is assumed to be known continuously along each chromosome. Real

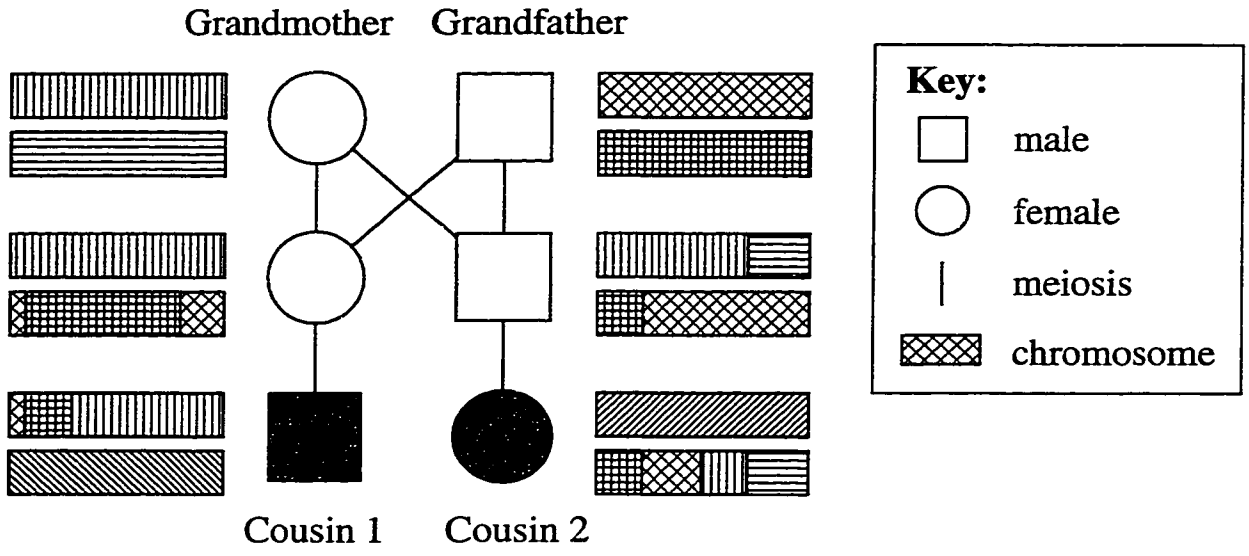

Figure 1.1: **Copies of a single chromosome in individuals on a cousins pedigree.** Each individual has a maternal and a paternal copy. At meiosis, the two chromosome copies combine into one copy to be passed down to the next generation.

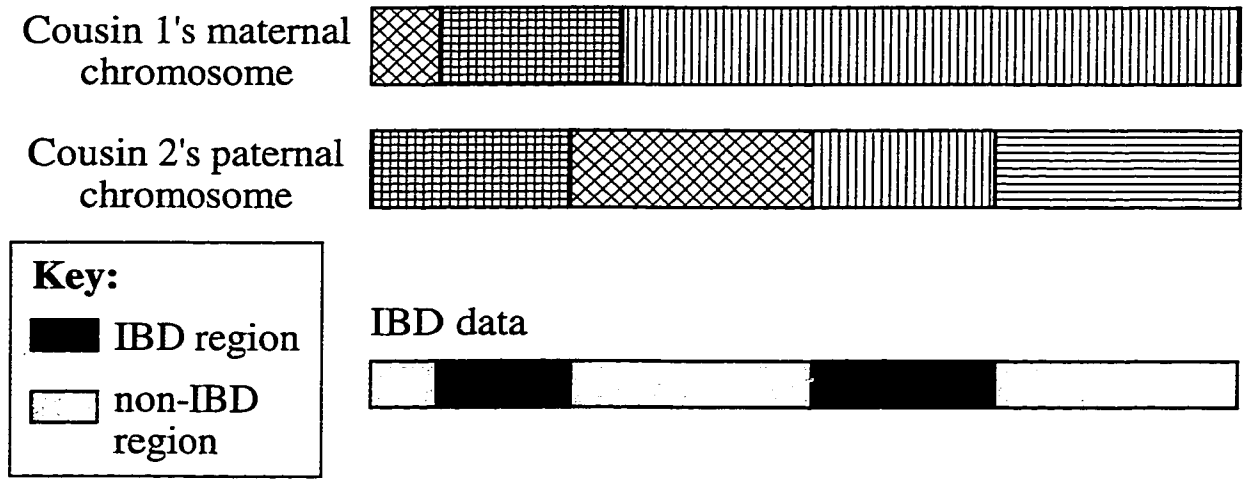

Figure 1.2: **Cousins' IBD data.** Comparison of the cousins' chromosomes reveals matching due to identity by descent.

data are discrete, with genetic markers (which can be thought of as short sections of DNA that come in several different forms, or *alleles*) located at discrete loci (points) along the genome. At each marker locus it is *identity by state* (IBS) status rather than IBD status that is observed. Two individuals are IBS at a marker locus if they share the same allele at that locus. The alleles may be identical by chance, since, for each locus, the total number of alleles carried in the population is finite (and often small). If the individuals are related, the alleles may be identical due to inheritance from a common ancestor in the pedigree defining their relationship (in which case they are IBD at the locus). Identity by state at large numbers of adjacent loci tends to only occur in related individuals, and only in regions of IBD. It is unlikely that a large number of alleles will be shared simply by chance, unless there is little variability in the population along that section of DNA.

As genetic markers become increasingly densely spaced, it will become easier to infer IBD status essentially continuously along each chromosome. By looking at information content of idealized continuous IBD data, we examine the limit of information content in discrete IBS data as the markers become infinitely dense, assuming sufficient polymorphism in the population. For this reason, we choose to investigate questions of information content and power with regard to idealized continuous data. In chapter 4 we will present a modification of the Monte Carlo algorithm for likelihood calculation that can be used to analyze discrete data.

1.2 The process of crossing-over

The DNA of an individual is arranged in pairs of homologous chromosomes, with one homologue inherited from each parent. At meiosis a gamete (egg or sperm cell) is produced which has single copies of each chromosome. Each of the gamete's chromosomes may be entirely of maternal or paternal origin, or may be a mixture with points of crossing-over between the two parental origins. For example, in figure 1.1,

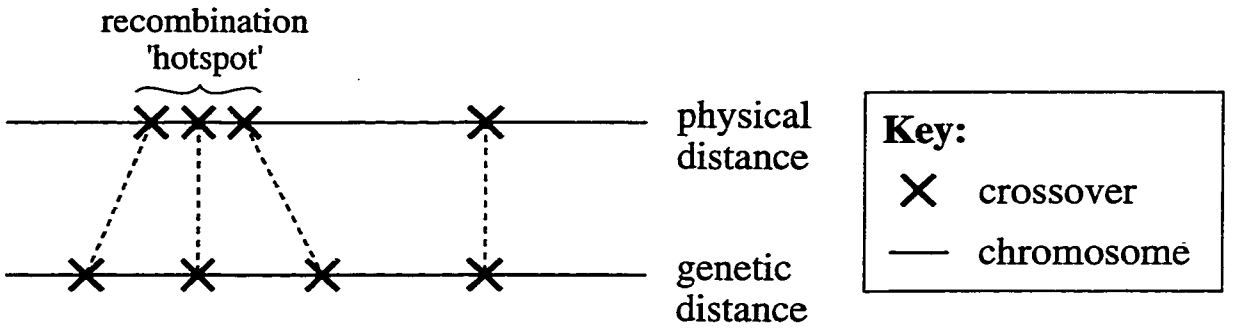

Figure 1.3: Physical versus genetic distance.

the mother of cousin 1 received one of her mother's (the grandmother's) chromosomes in its entirety while she received a mixture of her father's (the grandfather's) two chromosomes with two crossovers.

When measuring distance along the genome in molecular units, (i.e. base pairs), one finds that the rate of crossing-over is not at all constant along the genome. Instead there are 'hot spots' where crossovers occur more frequently than the genome-wide average, and 'cold spots' where crossovers occur less frequently. However, rather than model the change in rate along the genome, statistical geneticists locally rescale physical distance, to create a new measure of distance. The genetic distance between two loci on a chromosome is defined to be the expected number of crossovers between those loci in a gamete. Figure 1.3 illustrates the difference between the two measures. The unit of genetic distance is the Morgan, and genetic distances are estimated from observed recombination frequencies (for example, [12]).

Rates of crossing-over do differ between males and females and from one individual to another, but we will not consider such differences until section 3.7.

The crossing-over process at meiosis is a stochastic point process on each chromosome, with crossovers occurring at random locations along the chromosome. In addition, the parental origin (maternal or paternal) of the DNA passed down at the starting end (choice of "starting end" is arbitrary) of a chromosome is random. At

each crossover, the parental origin of inherited DNA switches.

A number of different models have been proposed for the crossing-over point process. In chapter 3 some of these models will be discussed. We will show how they can be incorporated into the Monte Carlo approach and present simulation results investigating power to distinguish between models. Most studies have looked at data from either single meioses or tetrads, for example Zhao et. al. [27], however IBD or IBS data may be more readily available.

The combination of all the crossover processes at each meiosis on the pedigree defining the relationship between two individuals determines their IBD process (see Donnelly [7]). This concept will be important in development of the Monte Carlo approach.

1.3 Relationship inference

Suppose that one has genetic IBS or IBD data for two individuals whose relationship is unknown (or possibly misreported). Those data can be used in making an inference about the true relationship between the individuals. Practical instances where such inference is useful occur in many fields.

In sib-pair studies, IBS data from pairs of brothers and sisters in which both siblings have a disease or trait of interest are used to find the locations of genes associated with the disease or trait. Chromosomal regions along which most of the pairs of siblings are IBS are candidate regions for the gene locations. Although the pairs are purported to be full brothers and sisters, most studies contain some pairs of half brothers and sisters (due to non-paternity) and perhaps some unrelated pairs (due to adoption or sample mix-ups). Such misreported pairs reduce power to find the genes of interest, so it is important to separate the half-sibs and unrelated pairs from the full-sibs. Boehnke and Cox [3] present an example in which evidence of linkage increases significantly on removal of misclassified pairs.

Some gene-mapping studies use genetic data collected from small nuclear families. If the families are part of a closed population in which individuals intermarry, there will be many relationships linking the nuclear families. These relationships may not be recorded but may be able to be inferred by analyzing identity by state data on pairs of individuals from different families. Including such relationships in the study will increase power to detect linkage. Wijsman and Amos [26] found that, for a given number of sampled individuals, extended pedigrees provide much more linkage information than do nuclear families.

In studies monitoring the well-being of endangered species of animals or plants, it is sometimes desirable to discover the relationships between individuals. Knowledge of population structure can be helpful in determining population management strategies. See for example the analysis of relatedness in California condors in Geyer et. al. [10].

In the case in which one wishes to distinguish between two possible relationships between a pair of individuals (for example, are they full-sibs or half-sibs?) based on identity by state data, the information in the data can be fully exploited only by calculating the full likelihood of each relationship given the observed data. In some cases analytic calculation is possible, but if one wishes to employ realistic models, such as those discussed in chapter 3, analytic calculation is generally impossible while the procedures presented in this thesis provide a method for calculating the likelihoods.

1.4 Notation

Let k be the number of relevant meioses (those that can affect IBD status) making up the relationship between the two individuals of interest. For example, for the cousins relationship shown in figure 1.1, the meioses from the father of cousin 1 to cousin 1 and from the mother of cousin 2 to cousin 2 do not affect IBD status (and are not shown in the diagram), whereas the four meioses from grandparents to their children and the two meioses from those children to their children (the cousins) are relevant,

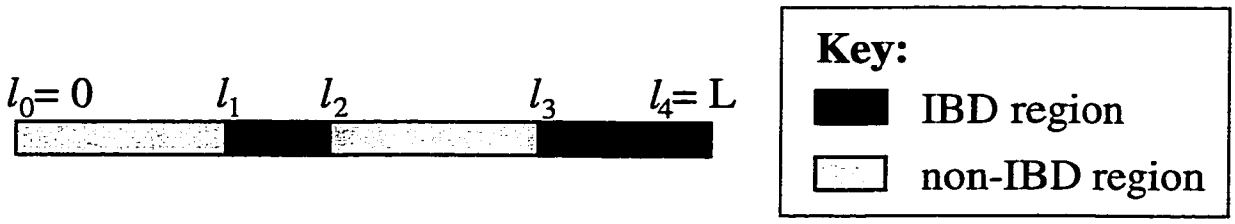

Figure 1.4: Locations of switches in IBD status.

so $k = 6$.

Let m be the number of IBD and non-IBD regions on a chromosome of continuous IBD data (so that there are $m - 1$ points of change in IBD status). Let $l_1 < l_2 < \ldots < l_{m-1}$ be the distances, measured in Morgans, from one end of the chromosome to the changes in IBD status. Also define $l_0 = 0$ and $l_m = L$, where L is the genetic distance from one end of the chromosome to the other. The example in figure 1.4 has $m = 4$ regions with switches in IBD status at locations l_1, l_2 and l_3.

Chapter 2

MONTE CARLO CALCULATION OF LIKELIHOODS

In this chapter we present a Monte Carlo algorithm for calculating likelihoods from continuous identity by descent data. The algorithm applies to any model for the crossover process, and in the next chapter we show how the method simplifies for some models. Simulation results presented later in the chapter investigate power to distinguish between relationships, in particular relationships that have the same expected proportion of identity by descent.

2.1 The combined crossover process

In order to develop the Monte Carlo approach, we first need to understand the processes underlying identity by descent.

Suppose we have identity by descent data on two related individuals, and a pedigree defining the relationship. For example, the pedigree might be the half-aunt-niece pedigree shown in figure 2.1, and the identity by descent data might consist of the comparison of the half-aunt's maternal chromosomes with the half-niece's paternal chromosomes.

A crossover process is a zero-one process with zeros representing maternal origin and ones paternal origin, with switches in origin occurring at the location of crossovers. For example, in figure 2.1, the crossover process for meiosis 1 starts with a zero indicating maternal origin, and then switches at the crossover to one indicating paternal origin. Compare this representation of the crossover process on meiosis 1 to the outcome of meiosis 1 in figure 2.2. The aunt (product of meiosis 1) receives first

her mother's maternal copy  , then the chromosome switches to her mother's paternal copy  .

Superimposing the crossover processes from all meioses on the pedigree gives the combined crossover process. This process records the location of crossovers on all meioses, and the indicators of parental origin for each meiosis at each location. At the location of a crossover, the indicator for the meiosis on which the crossover occurred switches, while the other indicators remain unchanged. Following the notation of Donnelly [7], the indicators of parental origin for the meioses may be represented as a vector with one element for each meiosis. For example in figure 2.1, the combined crossover process starts with maternal origin for meiosis 1 and paternal origin for meioses 2 and 3. This state is represented by the vector $(0, 1, 1)$. At the first crossover, which occurs on meiosis 3, the indicator for meiosis 3 switches from 1 (paternal) to 0 (maternal), and the new state of the process is represented by the vector $(0, 1, 0)$.

A subset of the possible vectors of parental origin correspond to identity by descent. For example, the states $(1, 1, 0)$ and $(0, 0, 0)$ are states of identity by descent for the half-aunt-niece relationship shown in figure 2.2, since for these states, the aunt's mother passes the same copy (either both paternal or both maternal) to each of her two children, and the niece's father passes down to her the copy he received from his mother. Thus the combined crossover process determines the identity by descent process. The reverse is not true: the identity by descent data do not determine the combined crossover process. So the combined crossover process, which is not observed, is a hidden process underlying the observed identity by descent data.

It will be helpful at times to consider the jump chain of the combined crossover process, which records only the sequence of states (parental origin vectors) visited by the combined crossover process and not the distances between crossovers.

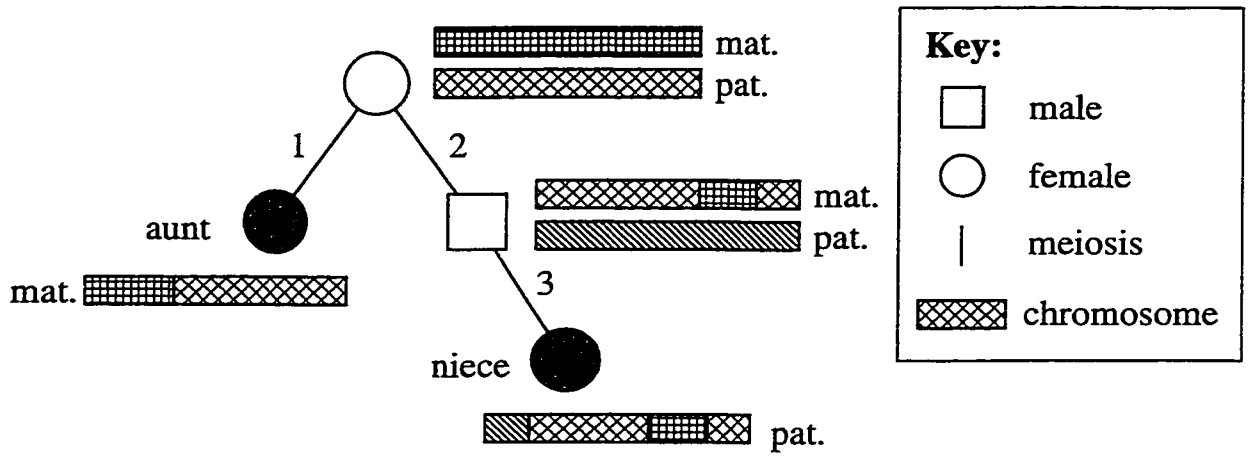

Figure 2.1: Crossover processes for meioses on a half-aunt-niece pedigree. Only relevant chromosomes are shown.

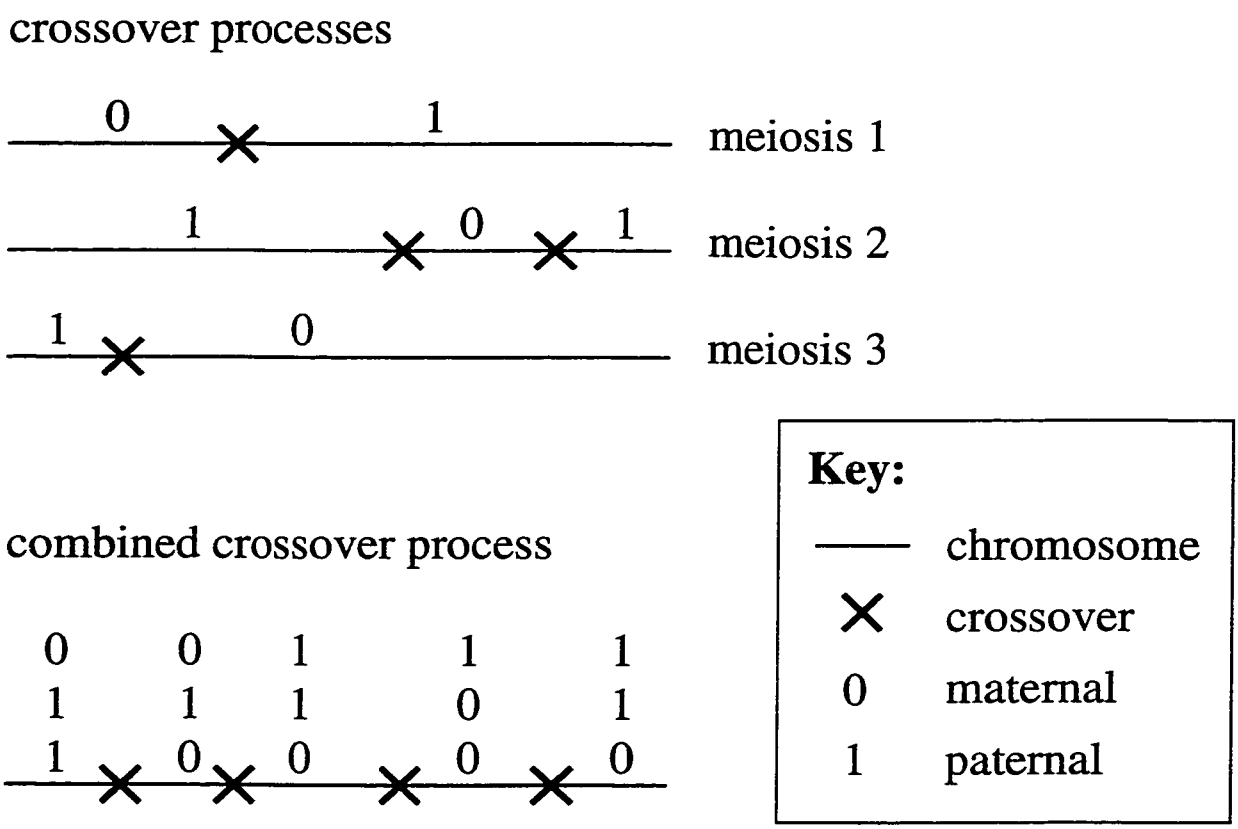

Figure 2.2: Combined crossover process for half-aunt-niece.

2.2 Monte Carlo algorithm

While we cannot, in general, calculate the probability of an IBD process realization analytically, we can calculate the probability of a crossover process realization for most conceivable crossing-over models. We will assume that crossover processes in different meioses are independent, so that calculation of the probability of a given combined crossover process realization is not difficult, provided the probabilities for the individual crossover processes have a simple form.

Since the combined crossover process is a hidden process determining the identity by descent process, it is sensible to approach the problem of calculating the probability of an IBD process realization by integrating over possible combined crossover processes:

$$P(\mathbf{I}) = \int P(\mathbf{I}|\mathbf{C})P(\mathbf{C})\, d\mathbf{C}$$

where $\mathbf{I}$ is any IBD process realization (i.e. IBD data) and $\mathbf{C}$ is any combined crossover process realization.

Now $P(\mathbf{I}|\mathbf{C})$ is zero unless $\mathbf{C}$ is consistent with $\mathbf{I}$ and is one if $\mathbf{C}$ is consistent with $\mathbf{I}$ since the combined crossover process determines the IBD process. Hence

$$P(\mathbf{I}) = \int 1\{\mathbf{C} \text{ consistent with } \mathbf{I}\}P(\mathbf{C})\, d\mathbf{C},$$

where $1\{\mathbf{C}$ consistent with $\mathbf{I}\}$ takes the value one if $\mathbf{C}$ is consistent with $\mathbf{I}$ and zero otherwise.

The integral is not itself tractable; we need a Monte Carlo approach to calculate it. Ideally we would like to take our Monte Carlo sample from a distribution proportional to $1\{\mathbf{C}$ consistent with $\mathbf{I}\}P(\mathbf{C})$, however we do not know how to sample directly from such a distribution, nor do we know the normalizing constant for this distribution. Instead we can sample from some other distribution $P^*(\mathbf{C})$. To do so, we can write

$$P(\mathbf{I}) = \int 1\{\mathbf{C} \text{ consistent with } \mathbf{I}\}\frac{P(\mathbf{C})}{P^*(\mathbf{C})}P^*(\mathbf{C})\, d\mathbf{C} \tag{2.1}$$

or

$$P(\mathbf{I}) = E_{P^*}(1\{\mathbf{C} \text{ consistent with } \mathbf{I}\}\frac{P(\mathbf{C})}{P^*(\mathbf{C})}) \tag{2.2}$$

provided P^* is non-zero on those combined crossover processes consistent with the IBD data. We will describe one reasonable choice for P^* in section 2.2.1.

If $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_n$ are sampled independently from P^*, a Monte Carlo estimate of $P(\mathbf{I})$ is

$$\hat{P}(\mathbf{I}) = \frac{1}{n}\sum_{i=1}^{n}\frac{P(\mathbf{C}_i)}{P^*(\mathbf{C}_i)}1\{\mathbf{C}_i \text{ consistent with } \mathbf{I}\}. \tag{2.3}$$

2.2.1 Choice of P^*

The sampling distribution P^* in equations 2.1, 2.2 and 2.3 above must be non-zero for all crossover processes that are consistent with the data. For importance sampling reasons, (so that the number of sampled realizations required to obtain an estimate with a given standard error is minimized) P^* should be as similar as possible to the distribution proportional to $1\{\mathbf{C} \text{ consistent with } \mathbf{I}\}P(\mathbf{C})$. In addition, P^* should be chosen so that it is both easy to sample from and simple to calculate.

The choice of P^* that will be used here is described first in terms of sampling (or simulating) a realization from P^* and the procedure is illustrated in figure 2.3. The hardest part in choosing a sampling distribution P^* is to achieve at least some amount of probability mass on the set of combined crossover processes consistent with the data. It is this task that drives our choice of P^*. Given data with starting IBD status s and switches of IBD status at locations $l_1, l_2, \ldots, l_{m-1}$, we will simulate first a jump chain (a sequence of states without locations) for the combined crossover process, and then we will simulate locations to complete a combined crossover process that is (usually) consistent with the data, as follows.

Simulate the jump chain starting with a state chosen at random from the IBD or non-IBD states of the state space according to s. Each of the states with IBD status s in the state space has probability $1/\#s$ of being chosen to be the starting state,

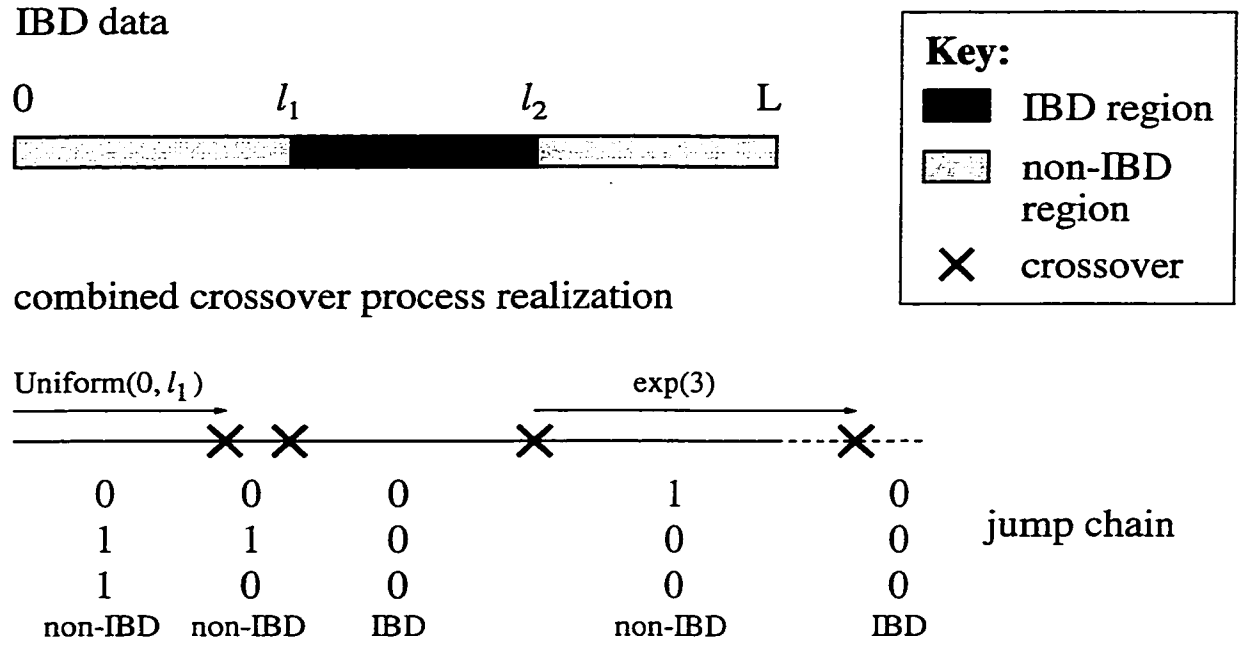

Figure 2.3: **Sampling under P^* for the half-aunt-niece relationship.** The jump chain is sampled first, then constrained crossover locations are set — the second and third crossovers must match l_1 and l_2 since they are points of change in IBD status. Unconstrained crossover locations in all but the last region are chosen uniformly within their bounds, so the location of the first crossover is chosen uniformly on $(0, l_1)$. Crossover locations for the last region are sampled from an exponential distribution with rate 3, since the half-aunt-niece relationship has $k = 3$ meioses.

where $\#s$ is the number of IBD states if s is IBD or the number of non-IBD states if s is non-IBD. The probability $1/\#s$ can also be written $2^{-k}/P(s)$ where $P(s)$ is defined as in Browning [6] to be the probability that the process starts with IBD status s, not conditional on the data. For example, suppose IBD data from the half-aunt-relationship shown in figure 2.2 start IBD (s =IBD). Then $\#s = 2$ since there are two IBD states (the states $(1,1,0)$ and $(0,0,0)$), $P(s) = 2/8$ since there are $2^k = 2^3 = 8$ states in total for this relationship. Hence the first state of the jump chain is chosen to be $(1,1,0)$ with probability $1/2$ and $(0,0,0)$ otherwise, in this case. On the other hand, if the IBD data started non-IBD (s =non-IBD) then $\#s = 6$ since there are $2^3 - 2 = 6$ non-IBD states for the half-aunt-niece relationship, and $P(s) = 6/8$. The first step of the chain is chosen in such a way that each of the 6 non-IBD states is equally likely to be chosen. For the example in figure 2.3, s =non-IBD, and the first state of the sampled jump chain is $(0,1,1)$ which is a non-IBD state.

Sample the jump chain step by step. At each step choose a meiosis at random (each of the k meioses has probability $1/k$ of being chosen) on which to place the next crossover (thus switching that meiosis indicator). For example, for the half-aunt-niece relationship, each of the three meiosis indicators is equally likely to be chosen. In the example in figure 2.3, the third meiosis is chosen at the first step, and the third meiosis indicator switches from 1 to 0, changing the state of the chain from $(0,1,1)$ to $(0,1,0)$. At the second step the second meiosis is chosen, and the state of the chain changes from $(0,1,0)$ to $(0,0,0)$.

Sampling of the chain continues until the chain has m complete IBD/non-IBD regions and the first state of region $m+1$, although not all of these steps may be needed once locations are added, as some of the steps may fall beyond the end of the chromosome. For the example in figure 2.3, the data have $m = 3$ regions, and the jump chain is sampled until the beginning of the fourth IBD/non-IBD region.

Let c_i be the number of steps corresponding to the ith region of the chain. For example in figure 2.3, $c_1 = 2$, $c_2 = 1$ and $c_3 = 1$.

Once the jump chain is sampled, we need to sample locations for the crossovers that will give a combined crossover process that is consistent with the data. The locations of crossovers that change the IBD status of the process are constrained to be placed at the locations of the corresponding switches in IBD status of the data. The remaining crossovers may be located anywhere on the chromosome, provided the order of steps of the chain is preserved. Within region i ($i = 1, \ldots, m-1$), choose the ordered locations of the $(c_i - 1)$ crossovers that are not determined by the data to be uniformly distributed within the region. A given set of crossover locations for the ith region has probability density $(c_i - 1)!/(l_i - l_{i-1})^{c_i - 1}$. For example in figure 2.3, the positions of the second and third crossovers are fixed by the data since they correspond to changes in IBD status. The position of the first crossover is sampled uniformly from the interval $(0, l_1)$.

The crossovers in the last (mth) region are constrained by the data only in so far as the location of the final sampled step of the chain (the first step of the $(m+1)$th region) must fall beyond the end of the chromosome. It is possible to devise schemes that ensure that the final sampled crossover does fall beyond the end of the chromosome (for example the scheme used in section 3.1), however they tend to make the calculations slightly more complicated which increases computer time per iteration. Instead, we will sample the locations for the final region in such a way that the calculations will be simple while the combined crossover process will not be inconsistent with the data too often. For this last region, we will simulate distances between adjacent crossovers from an exponential rate k distribution until the end of the chromosome is reached. If the end of the chromosome is beyond the location of the final sampled crossover, the combined crossover process is inconsistent with the data.

Let $\tilde{c}_m$ be the number of steps of the last region that fit within the length of the chromosome. For the example in figure 2.3, $\tilde{c}_m = 0$ as the sampled location for the first step of the third and last region of the process falls beyond the end of the

chromosome. Note that if $\tilde{c}_m \geq c_m$ the combined crossover process realization is not consistent with the data since a further IBD/non-IBD switch has occurred before the end of the chromosome.

Let $x_1, x_2, \ldots, x_{\tilde{c}_m}$ be the sampled exponential distances between crossovers of the final region that fit within the length of the chromosome. The ith sampled distance x_i contributes a density term ke^{-kx_i} to the probability. The probability that the $(\tilde{c}_m + 1)$th sampled length is greater than $L - l_{m-1} - \sum_{i=1}^{\tilde{c}_m} x_i$ and thus falls beyond the end of the chromosome is $\exp(-k[L - l_{m-1} - \sum_{i=1}^{\tilde{c}_m} x_i])$. So in total the sampled lengths for the crossovers in the last region of the process have probability $(\prod_{i=1}^{\tilde{c}_m} ke^{-kx_i}) \exp(-k[L - l_{m-1} - \sum_{i=1}^{\tilde{c}_m} x_i])$ which simplifies to $k^{\tilde{c}_m} e^{-k(L-l_{m-1})}$.

To simplify the notation in equation 2.4, let c be the total number of steps of the sampled combined crossover process that fit within the length of the chromosome. That is, $c = \tilde{c}_m + \sum_{i=1}^{m-1} c_i$. For the example in figure 2.3, $c = 3$ since $c_1 = 2$, $c_2 = 1$ and $\tilde{c}_3 = 0$.

The probability P^* described as a sampling distribution above is made up of the following parts: the choice of the first state of the jump chain has probability $2^{-k}/P(s)$; the choice of meiosis indicator to change at each step of the chain/process has probability $1/k$ at each of the c steps that fit within the length of the chromosome, so probability k^{-c} in all; and the choice of locations for the crossovers that aren't constrained by the data has probability $(c_i - 1)!/(l_i - l_{i-1})^{c_i - 1}$ for the ith non-end region and probability $k^{\tilde{c}_m} e^{-k(L-l_{m-1})}$ for the end region. Hence P^* can be written

$$P^*(\mathbf{C}) = 2^{-k} P(s)^{-1} k^{-c} \left(\prod_{i=1}^{m-1} \frac{(c_i - 1)!}{(l_i - l_{i-1})^{c_i - 1}} \right) k^{\tilde{c}_m} e^{-k(L-l_{m-1})}. \tag{2.4}$$

2.2.2 Modifications to the Monte Carlo algorithm

In chapter 3 we will see that the Monte Carlo algorithm can often be simplified, depending on the choice of model for the crossing-over process. In particular, sampling the locations of the combined crossover process is unnecessary if the locations do not

enter into the calculation of $P(\mathbf{C})$. In such cases it is sufficient to sample the jump chain only. Some simplifications require a slightly different choice of P^*, particularly with regard to the sampling of locations for the end region. For example, with such a modification to P^*, the algorithm described in this chapter simplifies to the Monte Carlo algorithm in Browning [6] for Haldane's model, as will be described in section 3.1.

2.3 Relationships with same expected IBD proportion

Borel [4] provides an example of relationship inference from IBD frequency data. In Borel's example, IBD frequencies from 100 pairs of individuals who were either all brothers or all cousins were used to distinguish between the two possible relationships. IBD frequency data may be used to distinguish between relationships with different expected proportions of IBD but will not help in distinguishing between relationships with the same expected proportion of IBD.

Thompson [25] noted that relationships with the same expected proportion of IBD genome may differ in expected patterns of IBD at two or more loci. The patterns thus provide information that can be used to help discriminate between relationships with the same expected proportion of IBD genome.

For example, suppose we have IBD or IBS data for a pair of individuals of unknown relationship and we want to distinguish between half-cousins once removed, second cousins and quadruple third cousins (shown in figure 2.4) as possibilities for the relationship between the individuals. These three relationships all have expected IBD proportion 1/16, so the proportion of IBD or IBS in the data cannot be used to distinguish between the relationships, but patterns of IBD or IBS differ for the three relationships. The average length of an IBD region, for example, is 1/5 Morgans for half-cousins once removed, 1/8 Morgans for second cousins and 1/30 Morgans for quadruple third cousins, so the length of IBD regions contains information about the

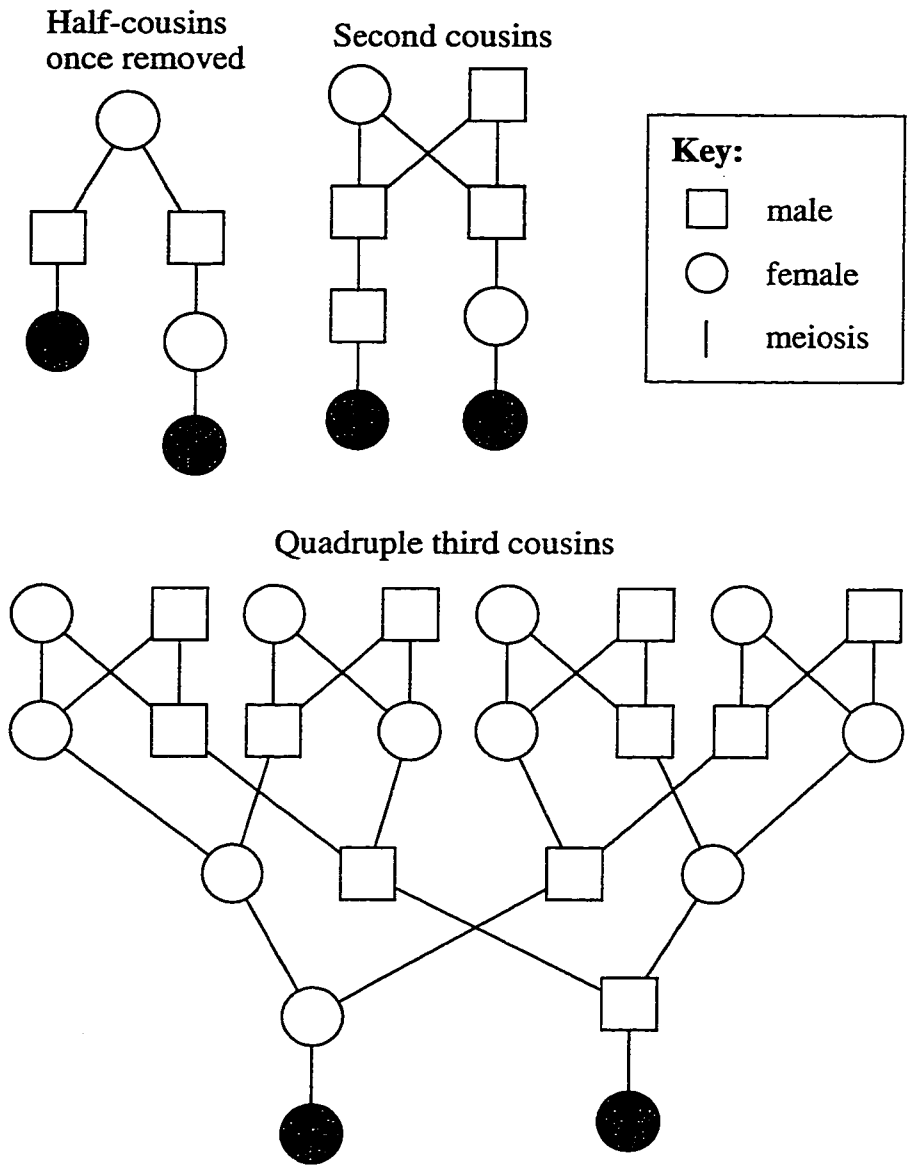

Figure 2.4: Three relationships with the same expected proportion of IBD genome.

relationship, as do other pattern features. The simulation study that follows examines the amount of information in IBD data for distinguishing between these three relationships, assuming Haldane's model for crossing-over.

As a proxy for information content, we ask the question "How many chromosomes of length 1 Morgan are needed to distinguish between a given pair of these relationships with 90% power at a 10% significance level?". That is, if one relationship is assigned to the null hypothesis and the other to the alternative hypothesis (because the chosen significance level equals one minus the chosen power, the result doesn't depend on which relationship is assigned to which hypothesis), how many chromosomes of data are needed so that the 90th percentile of the distribution of the likelihood ratio $L(\text{alternate})/L(\text{null})$ under the null hypothesis is equal to (or less than) the 10th percentile of the distribution of the same ratio under the alternative hypothesis.

Details of the simulation experiment used to answer this question follow in section 2.3.1. For half-cousins once removed versus quadruple third cousins we found that 35 to 40 chromosomes of length one Morgan are required. For half-cousins once removed versus second cousins approximately 250 chromosomes are required.

The length of the human genome is about 40 Morgans, and as shown in Browning [6] the relative lengths of the chromosomes do not have much effect on information content. Hence IBD data on the length of the human genome will generally be sufficient to distinguish between quadruple third cousins and half-cousins once removed, but there will not usually be enough information to distinguish between half-cousins once removed and second cousins.

Another case study, that of gametes from greatgrandparent-greatgrandchild (or half-sibs thrice removed) versus gametes from cousins (or second cousins), can be found in Browning [6].

2.3.1 Details of the simulation study

Ten thousand chromosomes of IBD data of length 1 Morgan from each relationship were simulated, and likelihoods of each relationship were calculated for each chromosome using 100,000 iterations of the Monte Carlo procedure. The actual Monte Carlo procedure used was that described in Browning [6], which as mentioned in section 2.2.2, is a simplification of the Monte Carlo algorithm presented in section 2.2. The simplification reduces computing time compared to the full algorithm in section 2.2.

Given a value N for the number of chromosomes in a data set, we can bootstrap the simulated data to find an estimate and confidence interval for the power to distinguish between the two relationships using IBD data from N chromosomes of length 1 Morgan. To perform the bootstrap, we sampled 1000 sets of N chromosomes of IBD data with replacement from the 10,000 chromosomes of IBD data. A log likelihood ratio was calculated for each set of N chromosomes of data, by adding together the logs of the individual chromosome likelihoods calculated with the Monte Carlo algorithm. The 1000 sets were used to estimate 90th percentile under the null distribution, and the proportion of the alternate distribution above that level (an estimate of power). The bootstrap procedure was repeated 10 times and the range and standard deviation of the ten power estimates was used to obtain an approximate 90% confidence interval for power.

If the upper limit of a 90% (say) confidence interval for power with N_1 chromosomes falls below 0.9 and the lower limit of a 90% confidence interval for power with N_2 chromosomes falls above 0.9, then we can be at least 90% confident that the number of chromosomes required to distinguish between the given pair of relationships with power 0.9 and significance level 0.1 is between N_1 and N_2.

For half-cousins once removed versus quadruple third cousins, with 40 chromosomes an approximate 90% confidence interval for power is (0.90–0.93) and with 35 chromosomes the approximate 90% interval is (0.86–0.90), so an approximate 90%

confidence interval for the number of chromosomes required to distinguish between these two relationships with power 0.9 and significance level 0.1 is (35–40).

For half-cousins once removed versus second cousins, approximate 90% confidence intervals for power are (0.86–0.89) with 220 chromosomes, (0.89–0.92) with 250 chromosomes and (0.91–0.94) with 280 chromosomes. Hence an estimate of the number of chromosomes of data required is 250, and an approximate 90% confidence interval for the number of chromosomes required is (220–280).

If an insufficient number of Monte Carlo iterations are used in calculating the likelihoods, power estimates will tend to be biased downwards because of the additional noise due to Monte Carlo error. As a check, the likelihoods on the simulated data were re-run with 1,000,000 Monte Carlo iterations for the half-cousins once removed versus second cousins case. There was no change in the results, indicating that 100,000 Monte Carlo iterations were sufficient in this case.

2.4 Other approaches to likelihood calculation

2.4.1 An exact method for discrete data with Haldane's model

Boehnke and Cox [3] present an exact method for calculating the probability of observed discrete data given a relationship. The approach assumes a Poisson process model for crossovers (Haldane's model) so that the combined crossover process is a Markov model. The method is based on Baum's [2] algorithm for hidden Markov models. Boehnke and Cox choose as the hidden process a variant on the IBD process, representing the state of the process as a vector with two indicators, one for IBD through the father and the other for IBD through the mother. This representation allows calculation of probabilities for sibling, half-sib and unrelated relationships. Although not discussed explicitly by Boehnke and Cox, it is natural to use the combined crossover process as the hidden process, allowing analysis of, in theory, any relationship. We will use this extension of Boehnke and Cox's method as part of the data

analysis in chapter 5.

If the relationship considered has k meioses, so that the combined crossover process has 2^k states, and if there are M markers on which the individuals are typed, the computation time for the algorithm is of order $O(M\ 2^{2k})$ so that the algorithm is in practice severely limited in the number of meioses that can be considered. To reduce this problem somewhat, symmetries can be exploited to partition the states into a smaller number of orbits (see Donnelly [7]). If the states are partitioned into B orbits, the computation time of the algorithm reduces to $O(M\ B^2)$. This is particularly useful when the number of meioses is large — for example the second cousins relationship has 8 meioses and hence 256 states, but the states can be partitioned into just 21 orbits [7].

Another limitation of this method is that it cannot deal with any type of crossover interference, as interference destroys the hidden Markov structure.

2.4.2 A numerical method for half-sib type relationships

One class of pedigrees involves symmetries that considerably simplify likelihood calculations. This is the class of relationships that Donnelly [7] refers to as 'half-sib type', in which the two individuals have only one common ancestor, and includes half-sibs, half-aunt-niece and so on. For each of these relationships the two states that correspond to IBD can be placed in the same 'orbit' (see [7] for details). A consequence is that points of entry and exit to IBD states form renewal points for the process. So the data can be split up with each IBD and non-IBD region treated separately. In addition, higher order coefficients of kinship for these relationships have a simple form which makes multi-locus IBD probability calculations possible.

Appendix A gives details of IBD probability calculations for these relationships. If we place more and more markers on a chromosome in a regular fashion and calculate the probability of observing the given sequence of IBD/non-IBD states at those markers then that probability, properly normalized, will converge to the likelihood.

Consider the half-sibs relationship and suppose that the length of a data chromosome is l and that there are m points of changeover between IBD and non-IBD regions on the chromosome. For such data the likelihood of the half-sibs relationship will be $\frac{1}{2}e^{-2l}2^m$. If $n+1$ points are placed at intervals of l/n along the chromosome and if n is large enough so that there is at least one point in each IBD and non-IBD region, then the probability of the sequence of IBD/non-IBD indicators is $\frac{1}{2}(\frac{1}{2}+\frac{1}{2}e^{-4l/n})^{n-m}(\frac{1}{2}-\frac{1}{2}e^{-4l/n})^m$. This probability multiplied by $(\frac{l}{n})^{-m}$ converges to the likelihood as $n \to \infty$.

Now consider another relationship R of half-sib type. Let $P_n(R)$ denote the probability of the sequence of IBD indicators at points $0, \frac{l}{n}, \ldots, l$ on the data chromosome given relationship R, and $L(R)$ denote the likelihood of R given the sequence of IBD indicators. As n increases to infinity, $P_n(R)/P_n(\text{half-sibs})$ will converge to the likelihood ratio $L(R)/L(\text{half-sibs})$. Hence,

$$L(R) = \lim_{n\to\infty}(\frac{l}{n})^{-m}P_n(R).$$

Using the algorithm described in the appendix for calculating $P_n(R)$ for R of half-sib type, a good approximation to the likelihood can be calculated quite fast. This algorithm assumes Haldane's model for crossing over and cannot be extended to more general models.

Chapter 3

MODELS FOR CROSSING-OVER

In this chapter we present some models for crossing-over and with simulation results examine how easy or difficult it is to compare these models using identity by descent data. For many of these models, the Monte Carlo procedure simplifies to some extent, and these simplifications are noted.

3.1 Haldane's model

Haldane [12] modeled crossing-over as a Poisson process along each chromosome, which assumes that the location of the next crossover after any point does not depend on the locations of crossovers before that point. This property of memorylessness is not realistic (for example [19]), but the model is popular as a first approximation.

Under Haldane's model, the probability of a given crossover process realization is constant, regardless of the locations of crossovers. As a result, sampled locations for combined crossover process realizations in the Monte Carlo procedure are irrelevant and the sampling of locations need not be carried out. With a certain choice of end-region strategy for P^* (different to that described in section 2.2.1) the Monte Carlo algorithm simplifies to that described in Browning [6], as we will show.

Consider first a single crossover process with starting parental origin a (maternal or paternal) and crossovers at locations $x_1, \ldots, x_{\tilde{m}-1}$ ($\tilde{m}$ is the number of sections into which the crossovers divide the chromosome and equals one if the process has no crossovers). The probability of the process is

$$P(a, \mathbf{x}) = \frac{1}{2} e^{-x_1} \left(\prod_{i=2}^{\tilde{m}-1} e^{-(x_i - x_{i-1})} \right) e^{-(L - x_{\tilde{m}-1})} \tag{3.1}$$

where L is the length of the chromosome. The factor of $1/2$ is the probability of a, the term e^{-x_1} is the probability density of the distance to the first crossover, the terms $e^{-(x_i - x_{i-1})}$ are probability densities for the distances between adjacent crossovers, and the term $e^{-(L - x_{\tilde{m}-1})}$ is the probability of no crossover after the $(\tilde{m}-1)$ th crossover. The probability (3.1) simplifies to

$$P(a, \mathbf{x}) = \frac{1}{2} e^{-L}.$$

Thus, for a relationship with k meioses, the probability of the combined crossover process $\mathbf{C}$ on a chromosome of length L is simply

$$P(\mathbf{C}) = 2^{-k} e^{-kL}. \tag{3.2}$$

Sampling of the jump chain of a combined crossover process will be the same as that described in section 2.2.1. This is also the same as the sampling of jump chains in Browning [6]. If the chain has c_i jumps corresponding to region i of the data $(i = 1, 2, \ldots, m)$ as in section 2.2.1, the probability of sampling the chain is

$$P^*(\text{jump chain}) = 2^{-k} P(s)^{-1} k^{-\sum_{i=1}^{m} c_i}$$

with $P(s)$ defined in section 2.2.1.

As noted above, it is not necessary to actually sample locations for the crossovers, however we do still need a sampling distribution for the locations so that P^* is defined and can be substituted into the estimate in equation 2.3. The distribution of locations of crossovers in all but the last IBD or non-IBD region can be the same as that in section 2.2.1. That is, constrained locations of crossovers are set and then other locations are sampled uniformly within the bounds on them. For these regions, the probability of a set of locations for the $c_i - 1$ unconstrained crossovers in the ith region is

$$P^*(\text{crossover locations in region } i) = (c_i - 1)!/(l_i - l_{i-1})^{c_i - 1}$$

as in section 2.2.1.

For the last IBD or non-IBD region, we note that under Haldane's model the distances between adjacent crossovers will be independent exponential random variables with rate k (superposition of Poisson processes), hence the sum of n distances will be gamma(n, k) with density

$$f(x|n) = \frac{x^{n-1}\, e^{-kx}\, k^n}{(n-1)!}.$$

This motivates our choice of sampling distribution for the locations of crossovers in the last region.

If our chain has c_m crossovers in the last region, we will first sample the sum of the distances as a gamma(c_m, k) random variable, conditional on the sum being greater than the length of the last region on the data. So if the last region of the data has length $L - l_{m-1}$, the sampled sum X has density

$$\begin{aligned} & P^*_X(x|c_m, X > L - l_{m-1}) \\ &= \frac{x^{c_m-1}\, e^{-kx}\, k^{c_m}}{(c_m-1)!} \Big/ \int_{y>L-l_{m-1}} \frac{y^{c_m-1}\, e^{-ky}\, k^{c_m}}{(c_m-1)!}\, dy \qquad (3.3) \\ &= \frac{x^{c_m-1}\, e^{-kx}\, k^{c_m}}{(c_m-1)!} \Big/ \sum_{j=1}^{c_m} \frac{(L-l_{m-1})^{j-1}\, e^{-k(L-l_{m-1})}\, k^{j-1}}{(j-1)!} \qquad (3.4) \end{aligned}$$

with line 3.4 following from line 3.3 by induction since

$$\int_{y>L-l_{m-1}} \frac{y^n\, e^{-ky}\, k^{n+1}}{n!}\, dy = \frac{(L-l_{m-1})^n\, e^{-ky}\, k^{n+1}}{n!} + \int_{y>L-l_{m-1}} \frac{y^{n-1}\, e^{-ky}\, k^n}{(n-1)!}\, dy\,.$$

Let $L^* = l_{m-1} + X$ be the location of the final crossover (or equivalently, the length of the sampled combined crossover process). Sample the ordered locations of the remaining $(c_m - 1)$ crossovers of the last region uniformly within the region between l_{m-1} and L^* — this sampling distribution has probability density $(c_m - 1)!/(L^* - l_{m-1})^{c_m-1}$.

The probability of a given set of locations for the crossovers in the last region is the probability density for the sampled sum X multiplied by the probability density for the sampled locations of the remaining $(c_m - 1)$ crossovers,

$$P^*(\text{crossover locations in region } m)$$

$$= \frac{(c_m-1)!}{(L^*-l_{m-1})^{c_m-1}} \frac{X^{c_m-1}\, e^{-kX}\, k^{c_m}}{(c_m-1)!} \Big/ \sum_{j=1}^{c_m} \frac{(L-l_{m-1})^{j-1}\, e^{-k(L-l_{m-1})}\, k^{j-1}}{(j-1)!} \quad (3.5)$$

$$= e^{-k(L^*-L)}\, k^{c_m} \Big/ \sum_{j=1}^{c_m} \frac{(L-l_{m-1})^{j-1}\, k^{j-1}}{(j-1)!} \quad (3.6)$$

with line 3.6 following from line 3.5 by substituting $X = L^* - l_{m-1}$ and simplifying.

Hence

$$P^*(\mathbf{C}) = 2^{-k} P(s)^{-1} k^{-\sum_{i=1}^{m} c_i} \left(\prod_{i=1}^{m-1} \frac{(c_i-1)!}{(l_i-l_{i-1})^{c_i-1}} \right) e^{-k(L^*-L)}\, k^{c_m} \Big/ \sum_{j=1}^{c_m} \frac{(L-l_{m-1})^{j-1}\, k^{j-1}}{(j-1)!}$$

$$= 2^{-k} P(s)^{-1} e^{-k(L^*-L)} \left(\prod_{i=1}^{m-1} \frac{(c_i-1)!}{k^{c_i}(l_i-l_{i-1})^{c_i-1}} \right) \Big/ \sum_{j=1}^{c_m} \frac{(L-l_{m-1})^{j-1}\, k^{j-1}}{(j-1)!} .$$

The length of the sampled combined crossover process $\mathbf{C}$ is L^*, so equation 3.2 tells us that

$$P(\mathbf{C}) = 2^{-k} e^{-kL^*}$$

and hence

$$\frac{P(\mathbf{C})}{P^*(\mathbf{C})} = P(s) e^{-kL} \left(\prod_{i=1}^{m-1} \frac{k^{c_i}(l_i-l_{i-1})^{c_i-1}}{(c_i-1)!} \right) \left(\sum_{j=1}^{c_m} \frac{(L-l_{m-1})^{j-1}\, k^{j-1}}{(j-1)!} \right) .$$

With this choice of P^*, $1\{\mathbf{C}_i$ consistent with $\mathbf{I}\}$ always equals 1, so that the estimate of equation 2.3 becomes

$$\frac{1}{n} \sum_{i=1}^{n} P(s) e^{-kL} \left(\prod_{i=1}^{m-1} \frac{k^{c_i}(l_i-l_{i-1})^{c_i-1}}{(c_i-1)!} \right) \left(\sum_{j=1}^{c_m} \frac{(L-l_{m-1})^{j-1}\, k^{j-1}}{(j-1)!} \right)$$

as in Browning [6] (see equations 5, 6 and 9 of [6], substituting $\mathbf{D} = \mathbf{I}$, $d_i = l_i - l_{i-1}$ and $l = L$). Hence with this choice of P^* the method is the same as that in [6] and the locations of sampled crossovers need not be sampled.

3.2 Chiasma process models

The physical event underlying crossing-over is the formation of chiasmata. At meiosis, each chromosome duplicates to form two sister chromatids and the two pairs of

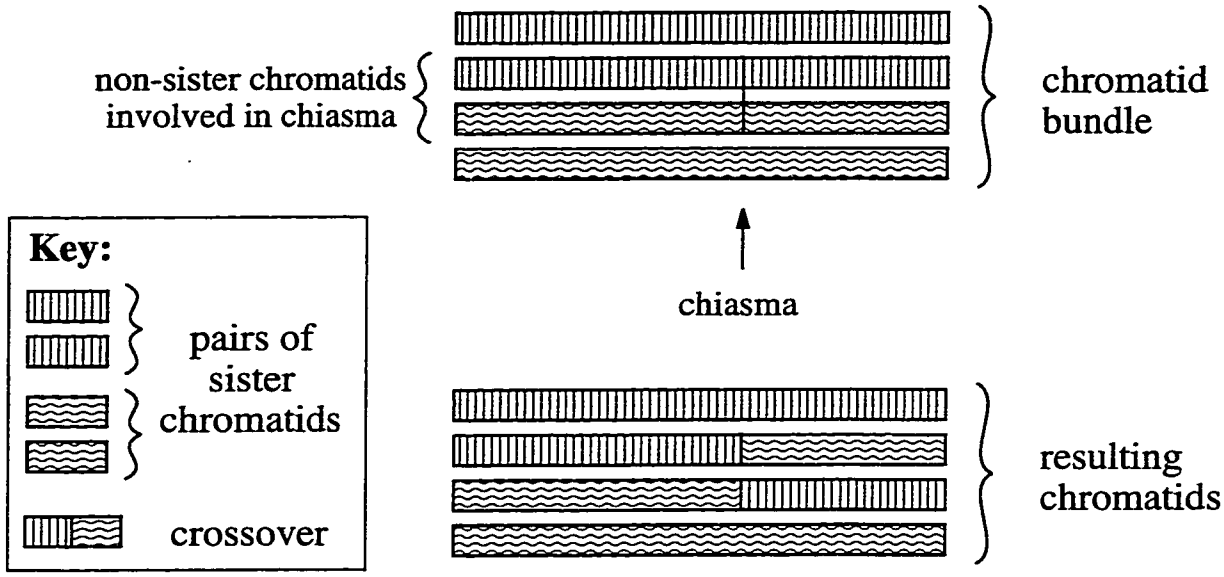

Figure 3.1: The process of chiasma formation.

homologous chromatids line up into a bundle of four. On this bundle, chiasmata occur, and each chiasma causes two non-sister chromatids to cross over. Figure 3.1 illustrates the process. Any one of the four resulting chromatids may be the one transmitted to the offspring.

It is common to assume no chromatid interference (NCI), so that each chromatid is involved in a crossover with probability one half at each chiasma independent of the outcome of other chiasmata on the bundle. This assumption seems to fit the available data (see [27]) and is convenient, although other models are possible and can be incorporated into the Monte Carlo approach.

A chiasma process model with either a model for chromatid interference or the assumption of NCI defines a corresponding crossover process model. In some cases, for example the truncated Poisson model in section 3.4, this crossover process model is easy to work with. However in most cases it is simpler to work directly with the chiasma process. There is no difficulty in modifying the Monte Carlo method to do this.

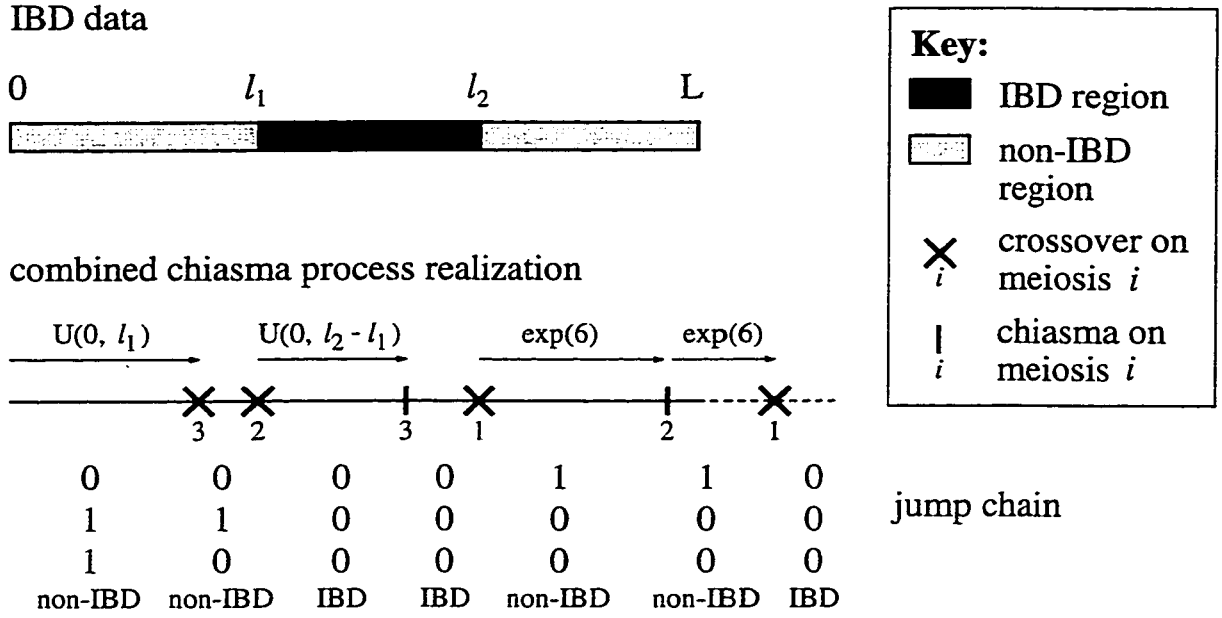

Figure 3.2: **Sampling of chiasma process under P^* for the half-aunt-niece relationship.** The jump chain is sampled first and records the sequence of chiasmata as well as whether each chiasma resolves as crossover in the child. Locations of constrained crossovers are set as in figure 2.3. Locations of unconstrained chiasmata are sampled uniformly within their bounds. Chiasma locations for the last region are sampled from an exponential rate 6 distribution since the half-aunt-niece relationship has $k = 3$ meiosis.

When working with the chiasma process directly we modify the sampling distribution P^* that was described in section 2.2.1. Figure 3.2 illustrates the process for the half-aunt-niece relationship. Sampling of the first state of the jump chain is as before. At each step of the chain one of the k meioses is chosen (each meiosis has equally probability of being selected) to receive a chiasma. With probability one half the chiasma resolves into a crossover and the meiosis indicator switches. Sampling of locations in regions other than the last region is as before, with locations sampled for chiasmata whether or not they resolved as crossovers. Locations for the chiasmata in the last region are sampled as exponential rate $2k$ random variables, since the rate of the combined chiasma process is $2k$. To calculate the probability of the sampled combined chiasma process, the individual chiasma processes for each meiosis can be separated out and the probability of each individual chiasma process calculated. This procedure will work for any chiasma process model under the assumption of no chromatid interference. It will also work for models of chromatid interference, although it will be inefficient if the chromatid interference is very strong.

3.3 Chi-square model

The chi-square model with integer parameter M models chiasmata along a chromosome as a renewal process with gamma$(M+1, 2(M+1))$ renewal density

$$f(x) = \frac{x^M e^{-2(M+1)x} (2M+2)^{M+1}}{M!}$$

(this is also the density of a scaled $\chi^2_{2(M+1)}$ random variable, hence the name of the model).

This model arises from considering chiasmata to be a special case of gene conversion, and from assuming that exactly M non-chiasma gene conversions occur between each adjacent pair of chiasmata, and that gene conversions (including those that resolve as chiasmata) occur as a Poisson process. Foss et. al. [8] give some background

into the gene conversion interpretation, however Foss and Stahl [9] reject the conversion interpretation because occurrence of conversions in a test on *S. cerevisiae* did not follow predictions of the model. Zhao et. al. [29] show that the model (if not the interpretation) fits available recombination data on a variety of organisms fairly well.

The similarity between this model with $M = 2$ and the Kosambi renewal model (described in section 3.5) is quite striking, as shown in figure 3.3. This is a special case of Zhao and Speed's [28] finding that 'stationary renewal processes give rise to most of the map functions in the literature ...[and] the interevent distributions of these renewal processes can all be approximated quite well by gamma distributions'.

One nice property of the chi-square model is that it exhibits almost complete interference over small distances. Figure 3.4 shows the cumulative distribution function for distances between adjacent crossovers for several values of M with Haldane's model (which is a special case of the chi-square model with $M = 0$) shown for comparison. The figure shows that for $M \geq 2$ there is almost complete interference over 5 centiMorgans (0.05 Morgans). For higher values of M the interference extends even further. There is evidence to suggest that interference may be almost complete over short distances in mammals, for example King et. al. [15] found very strong interference over a 14 centiMorgan region in mice.

The chi-square model can also be made more general by allowing variable numbers of non-chiasma conversions between adjacent pairs of chiasmata, as discussed in Appendix 1 of [8].

3.3.1 Monte Carlo calculation of likelihoods for the chi-square model

The chi-square model is a chiasma process model and fits into the Monte Carlo framework described in section 3.2. An alternative approach is to modify the method described in Browning [6] that is discussed in section 3.1.

Using the gene conversion interpretation of the chi-square model, we work with the combined *conversion* process (shown in figure 3.5) rather than the combined *crossover*

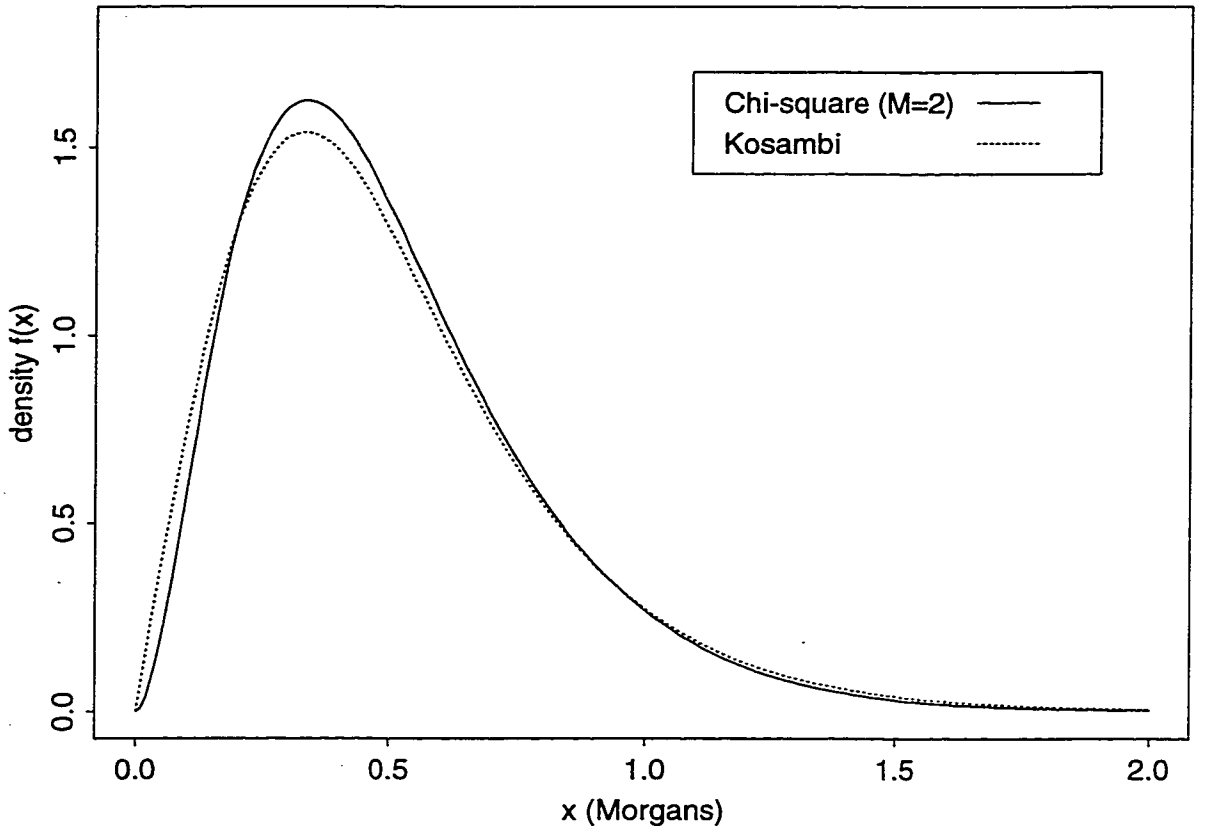

Figure 3.3: Comparison of renewal densities for the Kosambi and chi-square (M=2) models.

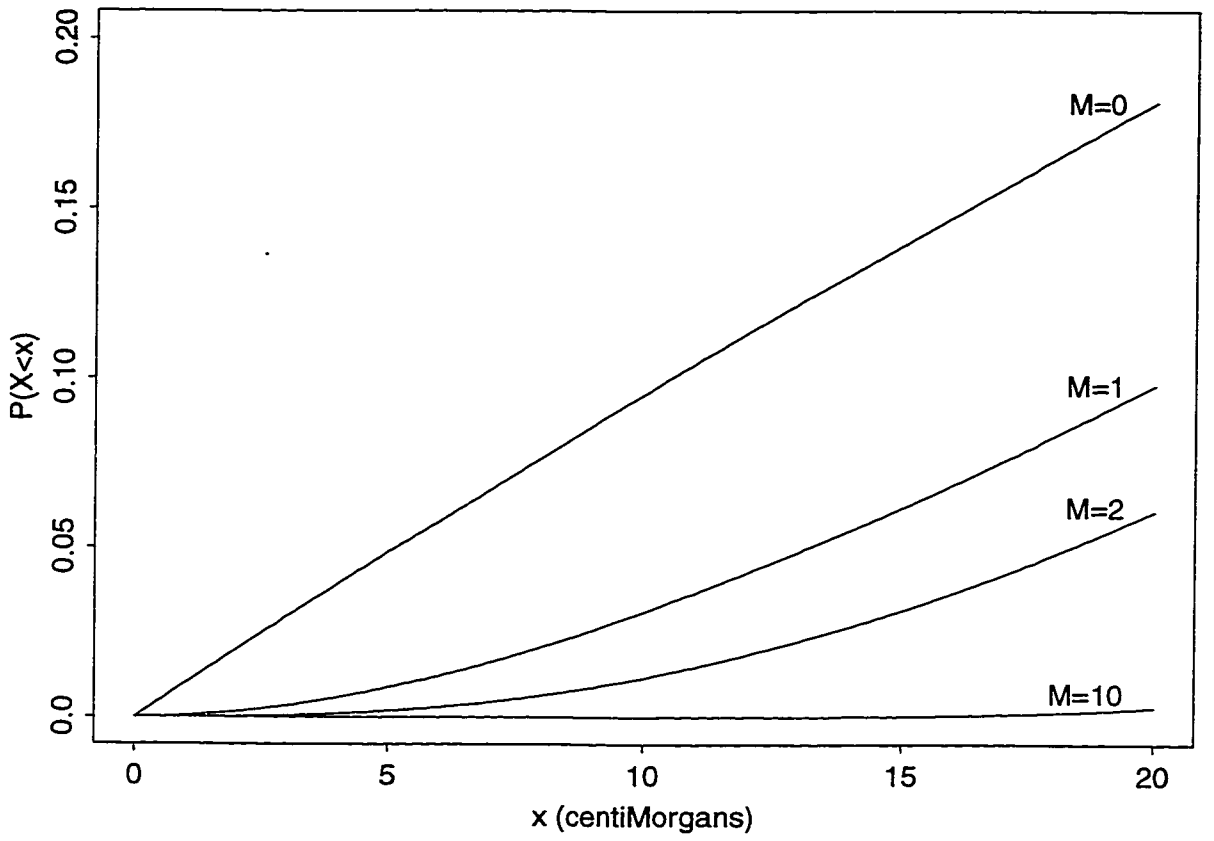

Figure 3.4: Cumulative distribution of distances between adjacent crossovers under chi-square models.

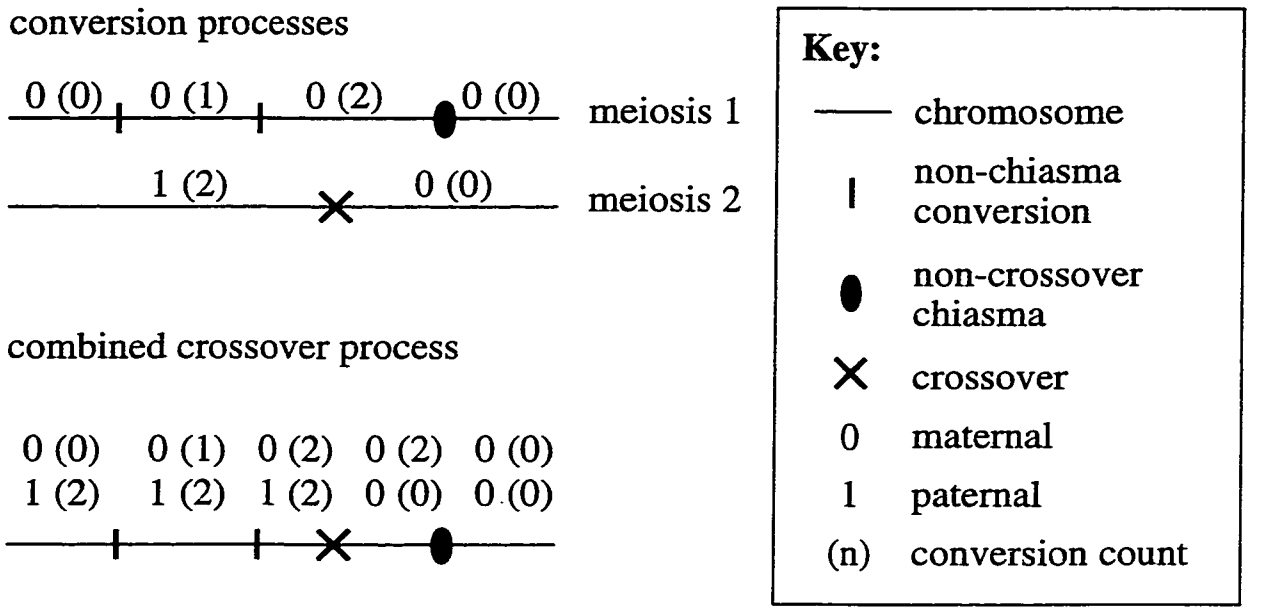

Figure 3.5: **The combined conversion process.** In this example there are $k = 2$ meioses, and the parameter for the chi-square model is $M = 2$. At the start of the chromosome, the conversion counter takes the value 0, 1 or 2 with equal probability. At each gene conversion, the conversion counter for that meiosis is incremented. When the conversion counter is incremented from 2 it returns to 0, and a chiasma occurs. The chiasma that occurred on meiosis 1 did not resolve as a crossover, whereas the chiasma that occurred on meiosis 2 did resolve as a crossover.

process. The state of the combined conversion process includes not only indicators of parental origin but also a count for each meiosis of the number of conversions since the last chiasma. At each step of the jump chain for this process, a meiosis is chosen to receive a gene conversion. The conversion count for the meiosis is incremented by one. If the new count equals $M + 1$ it is reset to zero and with probability one half a crossover occurs (assuming no chromatid interference so that resolution of chiasmata are independent).

Since gene conversion is assumed to occur as a Poisson process on each meiosis, and since meioses are independent, each meiosis is equally likely to be chosen for a conversion at each step, and the combined conversion process is a Markov process

with rate $2k(M+1)$. The probability of the IBD data (I) given a jump chain (C) is the same as that in [6] with the rate k there replace by $2k(M+1)$. Hence

$$f(\mathbf{I}|\mathbf{C}) = e^{-2k(M+1)L}\left(\prod_{i=1}^{m-1}\frac{2k(M+1)[2k(M+1)(l_i-l_{i-1})]^{c_i-1}}{(c_i-1)!}\right)\left(\sum_{j=0}^{c_m-1}\frac{[2k(M+1)(L-l_{m-1})]^j}{j!}\right).$$

3.4 Sturt and truncated Poisson models

The formation of at least one chiasma per chromatid bundle seems to be essential for correct disjunction of chromatids at meiosis. As a result, several chiasma process models have been proposed that ensure at least one chiasma per chromatid bundle. The Sturt [23] and truncated Poisson [17] models are both count-location models (see [17]), with the locations of chiasmata distributed uniformly and independently along the chromosome given N, the number of chiasmata.

For a chromosome of genetic length L, the Sturt model has the number N of chiasmata distributed as 1+Poisson($2L-1$) — that is, the model places one chiasma at a uniformly chosen random location on the chromatid bundle and then superimposes a Poisson chiasma process with mean $2L-1$ so that the overall number of chiasmata has mean $2L$ as required.

The truncated Poisson models has N following a Poisson distribution conditional on $N \geq 1$. The rate α of the distribution must be chosen such that the expected value of N is equal to $2L$, hence α solves $\alpha/2 = 1 - e^{-\alpha L}$. The probability distribution of N is, for $n \geq 1$,

$$\begin{aligned} P(N=n) &= \frac{\alpha^n L^n e^{-\alpha L}}{n!\,(1-e^{-\alpha L})} \\ &= \frac{2\,\alpha^{n-1} L^n e^{-\alpha L}}{n!} \end{aligned}$$

Speed [22] investigates properties of genetic mapping functions corresponding to stationary renewal chiasma processes. A mapping function $M(d)$ gives the probability

of recombination (an odd number of crossovers) in an interval of genetic length d. For the Sturt model [23],

$$M_S(d) = \frac{1}{2}\left[1 - (1 - \frac{d}{L})e^{-d(2L-1)/L}\right].$$

The mapping function for the truncated Poisson model is [17]

$$M_{tP}(d) = \frac{1}{2} - \frac{e^{-\alpha d} - e^{-\alpha L}}{2(1 - e^{-\alpha L})}.$$

Neither of these map functions satisfies Speed's condition A4. However since the total map length is necessarily finite (equal to L), this does not prove that the processes are not stationary renewal processes.

In appendix B we show, by direct calculation of probabilities for distances between chiasmata, that the truncated Poisson process is a renewal process while the Sturt process is not.

3.4.1 Calculation details for the truncated Poisson model

Assuming no chromatid interference, the crossover process of the truncated Poisson model is also a renewal process and has renewal density $g_2(x) = \frac{\alpha}{2}e^{-\alpha x/2}$, $0 < x < L$, and residual lifetime density $g_1(x) = e^{-\alpha x/2}$, $0 < x < L$. In this case it is more efficient (in the Monte Carlo method) to work with the crossover process than with the chiasma process, and we will demonstrate the details of the method.

Given locations of crossovers $x_1, x_2, \ldots, x_{c^*}$ along a chromosome from a single meiosis, the probability of the pattern is

$$g_1(x_1)g_2(x_2 - x_1)\ldots g_2(x_{c^*} - x_{c^*-1})(1 - G_2(L - x_{c^*})) = \left(\frac{\alpha}{2}\right)^{c^*-1} e^{-\alpha L/2}$$

provided $c^* \geq 1$, or

$$1 - G_1(L) = 1 - \frac{2}{\alpha}(1 - e^{-\alpha L/2})$$

when $c^* = 0$, where $G_1(x) = \frac{2}{\alpha}(1 - e^{-\alpha L/2})$ and $G_2(x) = 1 - e^{-\alpha L/2}$ are the cumulative distribution functions corresponding to the densities g_1 and g_2 respectively.

Consider a combined crossover process $\mathbf{C}$ with crossovers at locations $x_1, x_2, \ldots x_c$ (c is the number of crossovers that fall within the length of the chromosome, as defined in section 2.2.1). The probability of the process does not depend on the sequence of meioses on which the crossovers occur except through the number k_0 of meioses on which no crossovers occurred. The probability of the process is

$$\begin{aligned} P(\mathbf{C}) &= P(\text{starting state}) \prod_{i=1}^{k} P(\text{crossovers on meiosis i}) \\ &= 2^{-k} \left(\frac{\alpha}{2}\right)^{c-(k-k_0)} \exp(-\alpha(k-k_0)L/2) \left[1 - \frac{2}{\alpha}(1 - e^{-\alpha L/2})\right]^{k_0} \end{aligned}$$

where, as before, k is the number of meioses defining the relationship, and L is the length of the chromosome.

Hence, using the probability P^* defined in equation 2.4, the term $P(\mathbf{C})/P^*(\mathbf{C})$ in equation 2.3 equals

$$\begin{aligned} \frac{P(\mathbf{C})}{P^*(\mathbf{C})} &= \frac{2^{-k}\ (\alpha/2)^{c-(k-k_0)} \exp(-\alpha[k-k_0]L/2) \left[1 - (2/\alpha)(1 - e^{-\alpha L/2})\right]^{k_0}}{2^{-k} P(s)^{-1} k^{-c} \left(\prod_{i=1}^{m-1} \frac{(c_i-1)!}{(l_i - l_{i-1})^{c_i-1}}\right) k^{\tilde{c}_m} \exp(-k[L - l_{m-1}])} \\ &= P(s)(\alpha/2)^{c-(k-k_0)} \exp(-\alpha(k-k_0)L/2) \left[1 - (2/\alpha)(1 - e^{-\alpha L/2})\right]^{k_0} \times \\ &\quad \left(\prod_{i=1}^{m-1} \frac{k^{c_i}(l_i - l_{i-1})^{c_i-1}}{(c_i-1)!}\right) \exp(-k[L - l_{m-1}]) \end{aligned}$$

and

$$1\{\mathbf{C}_i \text{ consistent with } \mathbf{I}\} = 1\{\tilde{c}_m < c_m\}$$

where $\tilde{c}_m$ is the number of crossovers of the last region that fit within the length of the chromosome and c_m is the number of crossovers of the last region up to a further change in IBD status, as defined in section 2.2.1.

Note that the sampled crossover locations do not enter the expression for the term $P(\mathbf{C})/P^*(\mathbf{C})$, and hence, as for Haldane's model, the locations of crossovers need not be sampled.

3.5 Kosambi renewal model

The Kosambi map function [16]

$$M_K(d) = \tanh(2d)/2$$

is a map function that is commonly used when interference is being modeled. As Speed [22] points out, a map function is not itself sufficient to define a crossover or chiasma process. One solution provided by Karlin and Liberman [14] imposes rules for obtaining multi-locus probabilities from the two-locus probabilities given by map functions. However, for many map functions, including the Kosambi map function [21], this results in some negative probabilities. Karlin and Liberman call map functions for which negative probabilities result from their rules multi-locus infeasible.

A more satisfactory approach (discussed in [22]) is to derive a renewal process model that is consistent with the map function. This is not possible in every case but does work for the Kosambi map function [22]. We will call this model the Kosambi renewal model (given the map function, the renewal model is unique). Lange [17] shows that the renewal density for the Kosambi renewal chiasma process is

$$f_2(x) = 16\frac{e^{2x} - e^{-2x}}{(e^{2x} + e^{-2x})^3}$$

which can also be written as $f_2(x) = 4\sinh(2x)/\cosh(2x)^3$. The density of the distance to the first chiasma is

$$f_1(x) = 8/(e^{2x} + e^{-2x})^2$$

which can be written as $f_1(x) = 2/\cosh(2x)^2$.

The probability $P(\mathbf{C})$ of a given combined crossover process is straightforward to calculate but somewhat awkward to write out. There are no special simplifications to the Monte Carlo procedure for this model.

3.6 Model comparison

As genetic maps become more dense, it becomes both easier and more important to look at models for crossing-over. It is easier to look at models in the sense that the data contain more information about which models are better, although extracting that information is not necessarily easy. It is more important because assuming a wrong model will have a greater effect on probability calculations if the markers are densely spaced than if they are sparse.

As an example of the importance of assuming the correct model when calculating likelihoods of relationship using data with densely spaced markers, consider calculating the likelihood of a relationship in which very short non-IBD regions strongly suggest that the crossovers at each end of the region occurred on the same meiosis (this is the case, for example, for half-sib type relationships with at least three meioses). Suppose the true model has very strong positive interference over short distances. Under such a model, two crossovers on one meiosis within a short interval are almost impossible. Thus, if the dense data strongly suggest the presence of a very short non-IBD region, then the likelihood of that relationship is very low. If an incorrect model that has only weak interference is assumed, the calculated likelihood will be much higher than it should be. On the other hand, if the data are sparse, it won't be possible to detect short regions of non-IBD, and the likelihoods under the correct and incorrect models won't differ by much.

McPeek and Speed [19] examine the fit of various models to recombination data from Drosophila meioses. They look at recombinations between up to nine linked loci and assess the fit of the models to the data by comparing the observed numbers of each recombination pattern to those expected under the model, with the expected numbers being estimated by simulation. Their approach works well for the Drosophila data set but is limited in the number of loci that can be considered at once, since the number of recombination patterns that have to be considered grows exponentially

Table 3.1: **Number of chromosomes needed to distinguish between pairs of models.** Ranges give approximate 80% confidence intervals for the number of chromosomes required when chromosome length is 1 Morgan (above the diagonal) or 0.51 Morgans (below the diagonal). The chi-square model has parameter $M = 2$.

	Haldane	Chi-square	Trunc. Poisson	Kosambi
Haldane	—	34-46	230-370	45-57
Chi-square	70-90	—	46-60	1600-1900
Trunc. Poisson	20-26	40-60	—	60-82
Kosambi	80-120	1000-5000	41-55	—

with the number of loci.

The Monte Carlo approach presented in chapter 2 combined with the modifications for discrete data presented in chapter 4 provide a full likelihood method for comparing models using IBS data with any number of linked markers. In this section we will work only with continuous IBD data, but chapter 4 will discuss interference models in the context of discrete IBS data.

To examine the power to distinguish between the Haldane, chi-square with M=2, truncated Poisson and Kosambi models, we carried out a simulation study. With this study, we sought to answer the question: "With a significance level of 10%, about how many chromosomes of a given length of IBD data from half-sibs are needed to distinguish between two given models with 90% power". Note that since power equals one minus the significance level in this example, the results are not affected by which of the two models is chosen to represent the null hypothesis or alternate hypothesis. Table 3.1 presents the results as approximate 80% confidence intervals. Results below the diagonal of the table are for chromosomes of length 0.51 Morgans, while results above the diagonal are for chromosomes of length 1 Morgan.

Models, such as the truncated Poisson, which require at least one chiasma per

chromatid bundle, imply that chromosome lengths must be at least 0.5. So as a 'short' chromosome length we consider a chromosome of length 0.51, slightly greater than the minimum. We compare the results on these short chromosomes with chromosomes of length 1 Morgan which may be considered to be a more moderate length.

From the table, it can be seen that when comparing models other than the truncated Poisson, the number of chromosomes required to distinguish between the models approximately doubles when the chromosome length halves, so that the total length of DNA required remains approximately the same. This is similar to the result in Browning [6], in which it was seen that cutting chromosomes into smaller pieces (while keeping total genome length fixed) did not have a noticeable effect on power to distinguish between relationships.

Unlike the other models, the features of the truncated Poisson model change with chromosome length. If the chromosome length is close to 0.5 Morgans, most meioses will result in one and only one chiasma per chromatid bundle, which has the appearance of strong positive interference. As the chromosome length increases, the truncated Poisson model starts to look more like Haldane's model which has no interference. In the limit, as chromosome length increases to infinity, the truncated Poisson model is exactly Haldane's model. We see that with short chromosomes of length 0.51 Morgans it is very easy to distinguish between Haldane's model and the truncated Poisson model, with about 23 chromosomes required, while it is fairly difficult to distinguish between the two models when chromosome length is increased to 1 Morgan, with about 300 chromosomes required.

This aspect of the truncated Poisson model also carries over into comparison with the Kosambi and chi-square models. It is easier to distinguish between the Kosambi and truncated Poisson models when chromosomes are of length 0.51 Morgans (about 48 chromosomes required) rather than 1 Morgan (about 71 chromosomes required). When comparing the chi-square and truncated Poisson models it is unclear whether more chromosomes are required when chromosomes are of length 0.51 Morgans or 1

Morgan, but it is clear that the required total genetic length of the data is greater when the chromosomes are of length 1 Morgan than when they are of length 0.51 Morgans.

In section 3.3 we noted that the renewal densities for the chi-square model with $M = 2$ and the Kosambi model are very close. We see from the table that it is, unsurprisingly, very difficult to distinguish between these two models, with at least 1000 chromosomes required for chromosomes of lengths either 0.51 Morgans or 1 Morgan.

It should be noted that when comparing models we can include data from more than one pair of individuals in the analysis, so the amount of IBD data that can be analyzed is not restricted to the length of the genome. Hence, with sufficient data, it will be possible to distinguish between any given pair of these models.

3.6.1 Details of the simulation study

Ten thousand chromosomes (of length 0.51 Morgans and of length 1 Morgan) of half-sibs IBD data were simulated under each of the four models, and likelihoods under each of the four models were calculated using 100,000 iterations of the Monte Carlo procedure. The Monte Carlo algorithm described in Browning [6] and discussed in section 3.1 was used for calculating likelihoods under Haldane's model, and the extension of this method as discussed in section 3.3.1 was used for calculating likelihoods under the chi-square model. The simplification of the full Monte Carlo algorithm for the truncated Poisson model described in section 3.4.1 was used for calculating the likelihoods under that model. The full Monte Carlo algorithm as described in section 2.2 was used for calculating likelihoods under the Kosambi model.

To obtain approximate 80% confidence intervals for power, the same bootstrap procedure was carried out as for the study described in section 2.3.1, except that the bootstrap procedure was repeated 20 rather than 10 times for each combination of models and number of chromosomes.

3.7 Sex-specific rates of crossing-over

It is now evident that human males and females do not share the same genetic map (for example, Broman et. al. [5]). The order of loci along the autosomal chromosomes is the same for both sexes, but, when a female passes her DNA down to the next generation, the expected number of crossovers within a given region on a chromosome may differ from the corresponding expected number for a male. In particular, females tend to have higher crossover rates near the centromere relative to males, while males have a higher relative rate towards the telomeres. Also, the overall genetic length of the genome is longer for females than for males.

If we know the sex-specific maps relative to some base map, we can easily include sex-specific rates into the Monte Carlo method. The sampling distribution P^* may be the same as that described in section 2.2.1, with locations defined relative to the base-map. The sex-specific rates are incorporated into calculation of the actual probability P of the sampled processes. In calculating $P(\mathbf{C})$, the individual crossover processes (one for each meiosis) making up the combined crossover process $\mathbf{C}$ are separated, and the probability of each is calculated. At this point distances can be converted from the base map to the sex-specific map for each meiosis, and the calculations can then take into account the sex-specific map. Calculation of $P(\mathbf{C})$ depends on the model for crossing-over. As an example of calculation of $P(\mathbf{C})$, under Haldane's model (see section 3.1) the probability of each individual crossover process is $e^{-L}/2$ if L is the genetic length of the chromosome. So, if there are n_f female meioses and n_m male meioses, and if the lengths of the female and male maps are L_f and L_m respectively, then the probability of the combined crossover process is $2^{-k}e^{-n_f L_f}e^{-n_m L_m}$ where $k = n_f + n_m$ is the total number of meioses.

This method should work well as long as the base map is not too different from the male and female maps — i.e. the male and female maps must not differ by too much, and the base map should be chosen to be intermediate between the two sex-specific

maps (for example, the pedigree-average map of section C.2).

An alternative approach is described in appendix C. This approach modifies the method in Browning [6] and can be used with Haldane's model and the chi-square models.

These methods apply not only to sex-specific maps but also to individual-specific maps, however in humans it is unlikely that there will be many situations in which an individual-specific map is sufficiently well known to be used.

Chapter 4

DISCRETE DATA LIKELIHOODS

In this chapter we extend the Monte Carlo approach from idealized continuous data to real discrete data. Several issues that arise with real data must be considered. Compared to idealized continuous data, real discrete data are incomplete in two ways. First, identity by descent status is not known exactly at any point. Second, the discrete data have gaps between markers, so that even if IBD status were known at the marker locations, we would still not know exactly where switches in IBD status occur, and could even miss some small regions of IBD completely. An extra step is required in the Monte Carlo procedure to account for the incompleteness of discrete data.

4.1 Discrete data

As discussed in section 1.1, real data are discrete rather than continuous and provide IBS (identity by state) rather than IBD (identity by descent) status. Two individuals are IBS at a genetic locus if they share the same allele of the gene at that locus. Any given allele of a gene occurs with some frequency in the population, so that two unrelated individuals may share the same allele. Hence points of IBS suggest IBD but may also occur in regions of non-IBD. Also, small regions of IBD may fall between markers and hence cannot be observed in any way. Underlying any given IBS data there could be any of a great range of IBD processes. Figure 4.1 shows two possible underlying IBD processes for one IBS process. Of course, some of the possibilities will have much higher probabilities than others.

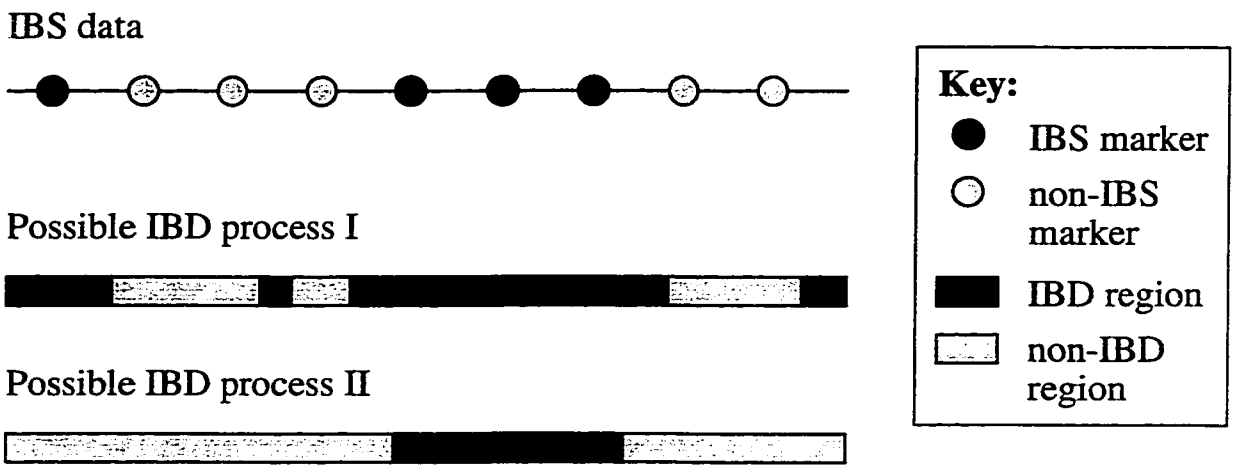

Figure 4.1: Two possible IBD processes underlying an example of discrete IBS data.

Given an IBD process realization, we can calculate the probability of the observed discrete IBS data. The IBS statuses at each marker are independent given the IBD process realization, and we can apply the probabilities in table 1 of [24] to each marker in turn. Note that the entry in row 2 (genotype pair of type 2) and column 6 (probability given one copy IBD) of the table should read $p_i^2 p_j$ rather than $p_i^3 p_j$.

4.2 Monte Carlo

There are two levels of hidden processes underlying discrete data: the unobserved continuous IBD process and below it the combined crossover process (or combined chiasma or conversion process). To calculate a likelihood, we need to integrate out both of these layers.

Let $\mathbf{D}$ be the discrete IBS data, let $\tilde{I}$ and $\tilde{J}$ be realizations of the continuous IBD data, and let P_A be a probability measure on continuous IBD data. An appropriate choice for P_A is discussed in section 4.2.1.

Integrating over IBD processes, and introducing P_A into the integral,

$$\begin{aligned} P(\mathbf{D}) &= \int P(\mathbf{D}|\tilde{I})P(\tilde{I})\,d\tilde{I} \\ &= \int P(\mathbf{D}|\tilde{I})P_A(\tilde{I})\frac{P(\tilde{I})}{P_A(\tilde{I})}\,d\tilde{I} \\ &= \left(\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})\,d\tilde{J}\right)\left(\int \frac{P(\tilde{I})}{P_A(\tilde{I})}\,\frac{P(\mathbf{D}|\tilde{I})P_A(\tilde{I})}{\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})\,d\tilde{J}}\,d\tilde{I}\right) \\ &\propto \int \frac{P(\tilde{I})}{P_A(\tilde{I})}\,\frac{P(\mathbf{D}|\tilde{I})P_A(\tilde{I})}{\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})\,d\tilde{J}}\,d\tilde{I} \end{aligned} \tag{4.1}$$

We can sample $\tilde{I}_1, \tilde{I}_2, \ldots \tilde{I}_n$ from

$$\frac{P(\mathbf{D}|\tilde{I})P_A(\tilde{I})}{\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})d\tilde{J}}$$

using a Metropolis-Hastings algorithm described in section 4.2.3. Let $\hat{P}_i(\tilde{I}_i)$ be the Monte Carlo likelihood estimate of equation 2.3 for some fixed number $\tilde{n}$ of iterations — whether one or many iterations are used, $\hat{P}_i(\tilde{I}_i)$ will be an unbiased estimate of $P(\tilde{I}_i)$. Then

$$\begin{aligned} \hat{P}(\mathbf{D}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{P}_i(\tilde{I}_i)}{P_A(\tilde{I}_i)}\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})d\tilde{J} \\ &\propto \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{P}_i(\tilde{I}_i)}{P_A(\tilde{I}_i)} \end{aligned} \tag{4.2}$$

is an unbiased and consistent estimator of $P(\mathbf{D})$.

The normalizing constant $\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})d\tilde{J}$ will be discussed in section 4.2.2 below.

4.2.1 Choice of P_A

Although P_A can essentially be any probability distribution, we would like it to be both easy to calculate and close to P. One choice that satisfies these requirements is the renewal approximation described in appendix D.2.2. Under the renewal approximation, lengths of regions are considered to be independent, that is, switches in

IBD status form renewal points for the process. We can then empirically estimate the probability distribution of the lengths of IBD and non-IBD regions from simulated data. As shown in appendix D the approximation can be very good.

4.2.2 The normalizing constant

When calculating a likelihood ratio, the normalizing constant $\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})d\tilde{J}$ will cancel provided the same choice of P_A is used for calculating both likelihoods. In this case P_A will not be so close to P for one or both models. A reasonable choice of P_A would be the renewal approximation for one of the models, or a mixture of the renewal approximation probabilities from the two models.

Provided that the number of meioses in the pedigree for the relationship is not too large, it is often most efficient to use the extension to Boehnke and Cox's [3] method described in section 2.4.1 when appropriate. If a likelihood is required for a model other than Haldane's, and if the number of meioses is not too large, a combination of the Monte Carlo method and the extension to Boehnke and Cox's method can be used. First, the Monte Carlo method can be used to calculate the likelihood ratio of the model of interest versus Haldane's model given the relationship. Then the extension to Boehnke and Cox's method can be used to calculate the likelihood of Haldane's model given the relationship. The likelihood of interest is found by multiplying the likelihood ratio by the likelihood of Haldane's model.

If the number of meioses for the relationship of interest is too large to be used with the extension to Boehnke and Cox's method, there are several possible approaches to calculating likelihoods. Firstly, the normalizing constant can be estimated directly by Monte Carlo, by simulating from P_A (or from some other probability measure on continuous IBD processes). For example, if $\tilde{J}_1, \tilde{J}_2, \ldots, \tilde{J}_n$ are sampled from P_A (which is straightforward to do — just sample region lengths from the empirical distributions until the chromosome length is reached), an estimate of $\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})d\tilde{J}$ is $\sum_{i=1}^{n} P(\mathbf{D}|\tilde{J}_i)/n$.

Secondly, one can first estimate the likelihood ratio of the model of interest versus a simpler model, with Haldane's model and a small pedigree. Then one can calculate the likelihood of the simple model using the extension to Boehnke and Cox's approach, and multiply the likelihood ratio by the simple model likelihood to get the likelihood of the model of interest. For example, suppose one wanted to calculate the likelihood of quadruple third cousins (see figure 2.4) under the chi-square model. One could calculate the likelihood ratio L(quadruple third cousins, chi-square model)/L(second cousins, Haldane's model) using the MCMC approach, with P_A based on a mixture of empirical region length distributions from the second cousin and quadruple third cousin relationships, and then use Boehnke and Cox's approach to calculate L(second cousins, Haldane's model).

4.2.3 Sampling realizations of $\tilde{I}$

Markov chain Monte Carlo (MCMC) can be used to sample from $\frac{P(\mathbf{D}|\tilde{I})P_A(\tilde{I})}{\int P(\mathbf{D}|\tilde{J})P_A(\tilde{J})d\tilde{J}}$ since both $P(\mathbf{D}|\tilde{I})$ and $P_A(\tilde{I})$ can be calculated easily and the normalizing constant does not need to be known for MCMC sampling. We will describe a Metropolis-Hastings [13] approach. The method is actually an example of reversible jump Markov chain Monte Carlo [11], since in adding or deleting IBD and non-IBD regions the dimension of the parameter space changes.

Proposals

Two types of proposal will be made — proposals to add or delete an IBD or non-IBD region, and proposals to move a switch in IBD status. The proposals are constrained by the fact that an IBD region should not cover a non-IBS marker. Figures 4.2 to 4.6 illustrate the proposals described below.

When the current IBD process is either entirely IBD or entirely non-IBD so that it has only one region, there is no switch in IBD status to move, and no IBD or non-IBD region to delete, so the proposal must be to add a region. On the other

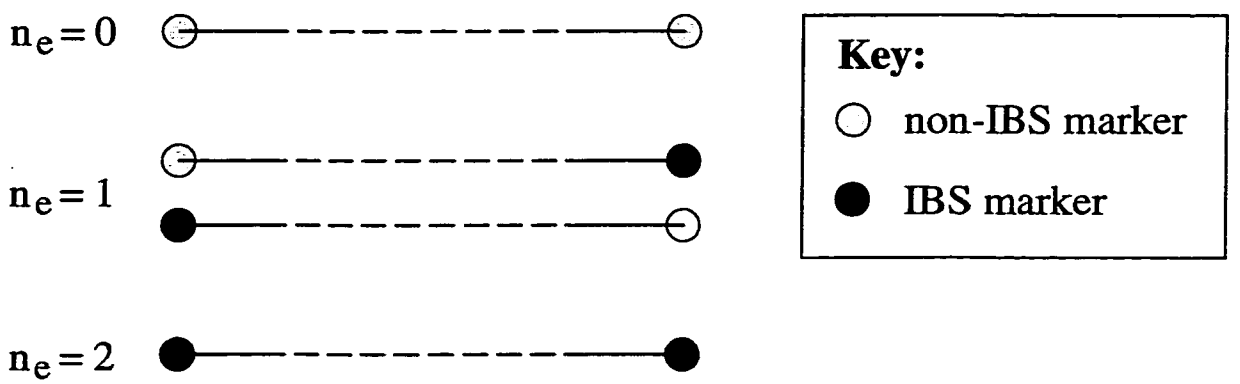

Figure 4.2: The number, n_e, of chromosome ends that can be both IBD and non-IBD is the number of chromosome ends that do not have a non-IBS marker at their extremity.

hand, if there is more than one IBD/non-IBD region in the current IBD process, we need to define some probabilities that will be used to select among the choices of add, move and delete. Set p_{move} between zero and one to be the probability of proposing a switch move (rather than add or delete) when there is more than one region. The probabilities of proposing to add or delete a region will each be $(1 - p_{\text{move}})/2$. We need to consider two types of region addition — adding a region that becomes the start or end region of the process is one type, and adding a mid region is the other. If the IBS data have a non-IBS marker right at one end of the chromosome, that end of the chromosome cannot be IBD, whereas if there is no marker at the end of the chromosome or if there is a marker and it is IBS, the end may be either IBD or non-IBD. Let $n_e \in \{0, 1, 2\}$ be the number of ends of the chromosome that do not have a non-IBS marker, as in figure 4.2. Set p_{end} between zero and one half and let $n_e\, p_{\text{end}}$ be the probability of adding a start or end region if the proposal is to add a region.

Let m^* be the number of switches in IBD status in the current IBD process, as illustrated in figure 4.3. When proposing a switch move, choose a switch uniformly at random from the m^* switches. Choose to move the switch left or right each with

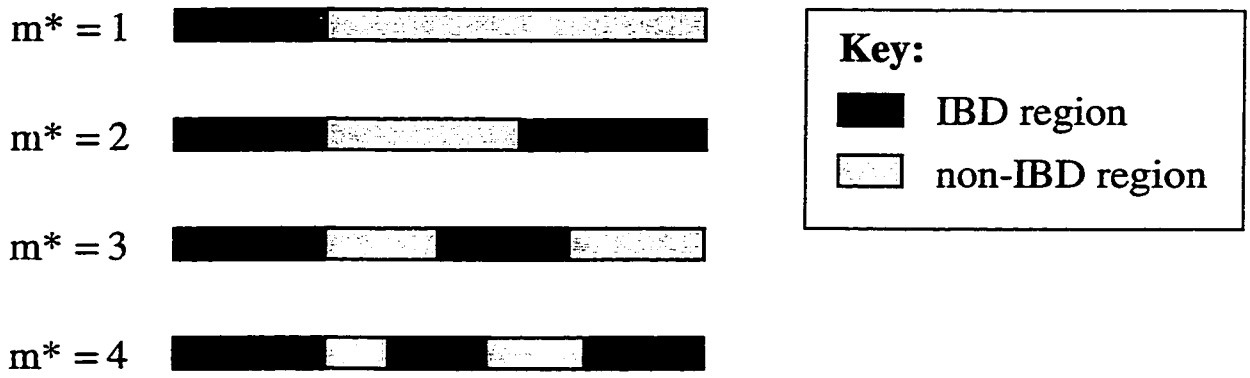

Figure 4.3: The variable m^* counts the number of switches in IBD status on an IBD process realization. This is equal to the number of regions minus one.

probability one half. The amount that the switch can move in that direction is constrained by the IBS data and by the current IBD process. The switch move must not result in IBD over a non-IBS marker, and we do not want to allow the switch move to change the number of IBD or non-IBD regions of the current IBD process, which would happen if we moved the switch up to or over another switch or up to the start or end of the chromosome. Let d^* be the maximum amount of movement possible in the chosen direction, constrained by the next switch, the start or end of the chromosome, and, if the switch move will increase the amount of IBD, the next non-IBS marker. It is reasonable to move the switch any distance between 0 and d^* in the chosen direction and we will sample the distance to move the switch from a uniform distribution over $(0, d^*)$.

When proposing to add an IBD or non-IBD region, choose to add an end region with probability $n_e\, p_{\text{end}}$. If $n_e = 2$, choose one of the two possible ends with equal probability. Let d^{**} be the maximum distance that the new region can extend towards the other chromosome end, taking into account non-IBS markers if the region being added is IBD, and existing switches in IBD status.

If proposing to add a mid region, choose a center point for the new region uniformly

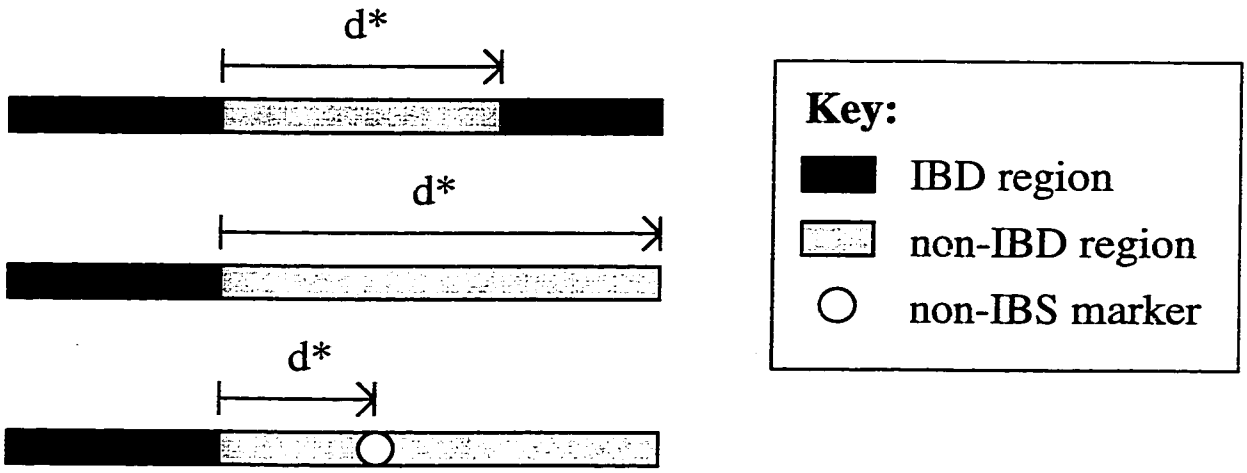

Figure 4.4: The variable d^* is the distance from the switch that will be moved to the next switch in IBD status, chromosome end or incompatible marker, in the proposed direction of the move.

at random along the chromosome length L. From the center point, investigate the range of distances that the new region can extend in either direction. Let d^{**} be the lesser of the maximum distances that the new region can extend to the right and to the left — taking into account the non-IBS markers if the region being added is IBD, the neighboring switches, and the chromosome ends, as in figure 4.5.

When proposing to delete an IBD or non-IBD region, determine which of the existing regions may be deleted and choose one at random to delete. IBD regions can always be removed but non-IBD regions can not be removed if they cover non-IBS markers. Let n^* be the number of regions for which removal is permitted by the IBS data, as illustrated in figure 4.6. Note that n^* will be at least one because if the IBD realization has two or more regions, at least one region will be an IBD region.

Acceptance probabilities

The Metropolis-Hastings acceptance probability is

$$a = \min\left\{1, \frac{\pi(\tilde{I}_2)q(\tilde{I}_2, \tilde{I}_1)}{\pi(\tilde{I}_1)q(\tilde{I}_1, \tilde{I}_2)}\right\} \tag{4.3}$$

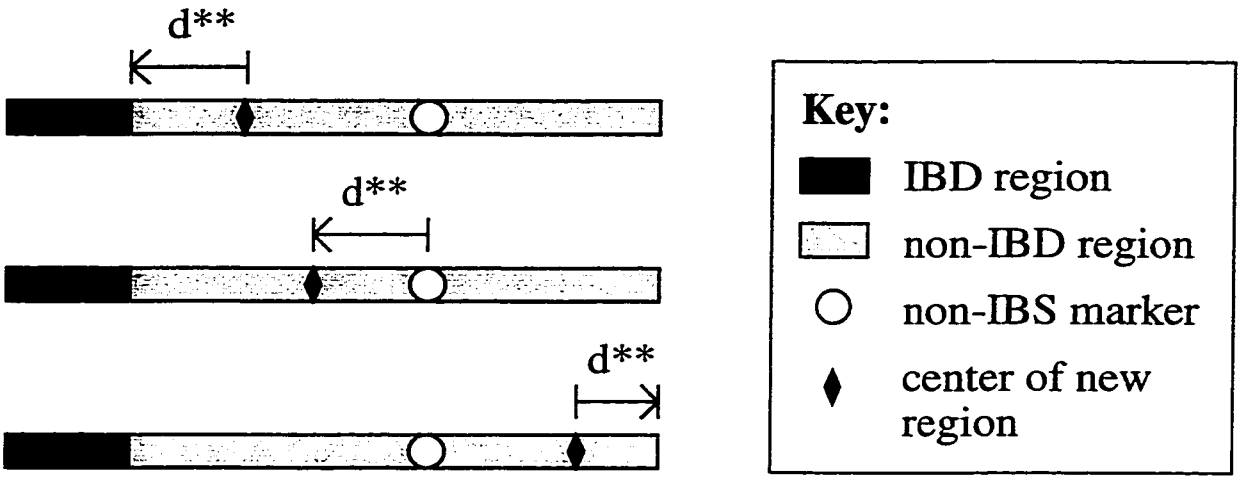

Figure 4.5: The variable d^{**} is the distance from the chosen center point of the new region to the nearest switch in IBD status, chromosome end or incompatible marker.

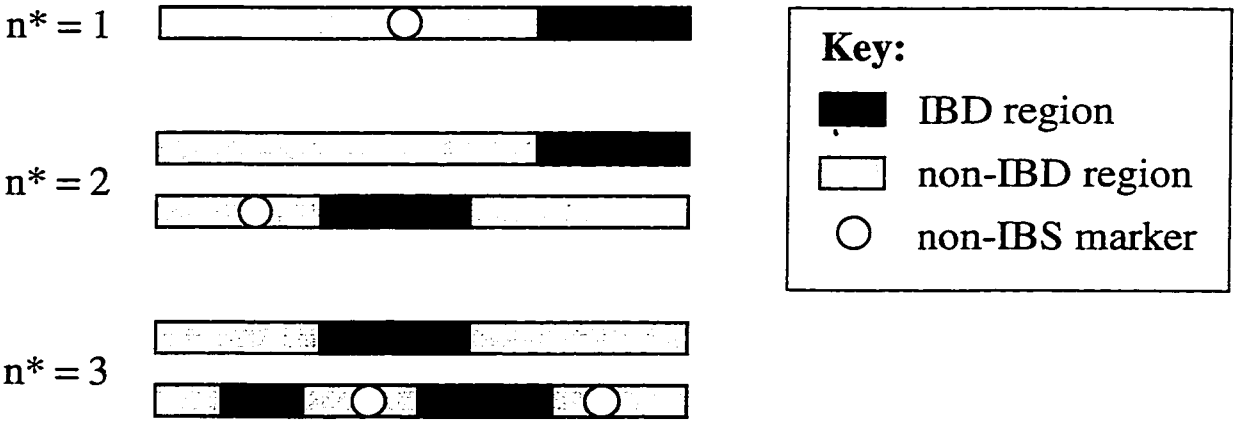

Figure 4.6: The variable n^* is the number of regions on an IBD process realization for which removal is permitted by the IBS data. This is equal to the number of IBD regions plus the number of non-IBD regions not covering non-IBS markers.

where $\tilde{I}_1$ is the current IBD process, $\tilde{I}_2$ is the proposed IBD process,

$$\pi(\tilde{I}_2)/\pi(\tilde{I}_1) = \frac{P(\mathbf{D}|\tilde{I}_2)P_A(\tilde{I}_2)}{P(\mathbf{D}|\tilde{I}_1)P_A(\tilde{I}_1)}$$

is the ratio of the probabilities of the proposed process to the current process and the q-ratio $q(\tilde{I}_2, \tilde{I}_1)/q(\tilde{I}_1, \tilde{I}_2)$ is the ratio of the probability of proposing $\tilde{I}_1$ from $\tilde{I}_2$ to the probability of proposing $\tilde{I}_2$ from $\tilde{I}_1$.

Note that for markers for which the IBD status is the same in both the current and proposed realizations, the probability terms in $\frac{P(\mathbf{D}|\tilde{I}_2)}{P(\mathbf{D}|\tilde{I}_1)}$ cancel, so that the terms only need to be calculated for markers for which the IBD status differs in the two realizations.

In what follows, we will describe calculation of the q-ratios. We will need to take into account 'dimension-matching' [11] when proposing to add or delete regions.

First we will consider a proposal to move a switch in IBD status. Since we need to consider both the proposed and reverse move we will use subscripts 1 and 2 to denote the proposal and reverse respectively. Hence for a proposed move right, d_1^* is the maximum move right from the current switch location, and d_2^* is the maximum move left from the proposed switch location.

Figure 4.7 shows d_1^* and d_2^* for an example move proposal. The probability density for moving a switch is $q(\tilde{I}_1, \tilde{I}_2) = \frac{1}{2}p_{\text{move}}/(d_1^* m^*)$, and the probability density for the reverse move is $q(\tilde{I}_2, \tilde{I}_1) = \frac{1}{2}p_{\text{move}}/(d_2^* m^*)$ where d_1^* is the maximum move permitted in the direction of proposed move, d_2^* is the maximum move permitted back in the reverse direction once the move has been made, and m^* is the number of switches in the current realization.

Now consider a region add. If a non-end region addition is being proposed, let $\ell_{1\bullet}$ and $\ell_{2\bullet}$ be the locations of the ends of of the new proposed region (see figure 4.8). The center $u_1 = (\ell_{1\bullet} + \ell_{2\bullet})/2$ of the new region is sampled uniformly from $(0, L)$, where L is the length of the chromosome. If d_1^{**} is the maximum radius of an added region with center at u_1, the radius of the new region, $u_2 = (\ell_{2\bullet} - \ell_{1\bullet})/2$, is sampled

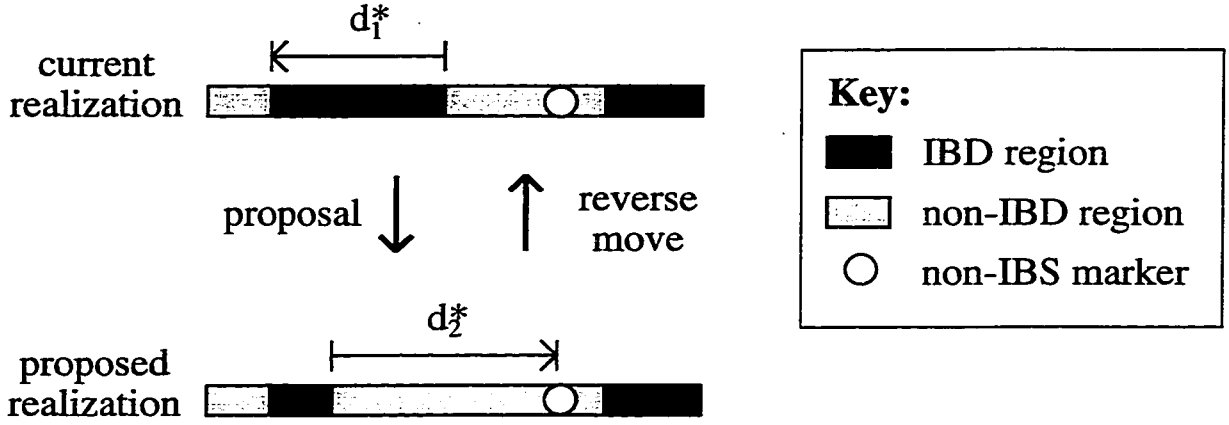

Figure 4.7: Move proposal and reverse move. In this example, the proposal is to move the second switch left. The maximum distance for this move is d_1^*. The actual amount of the move is chosen and the reverse move is considered for the q-ratio. The maximum amount that the second switch can move back to the right is d_2^*.

uniformly from $(0, d_1^{**})$. The probability density for proposing to add a mid-region with center u_1 and radius u_2 is

$$q(\tilde{I}_1, (\tilde{I}_1, u_1, u_2)) = \frac{1}{2}(1 - p_{\text{move}})(1 - n_e p_{\text{end}})/(d_1^{**} L) \tag{4.4}$$

if the current realization has more than one region, or

$$q(\tilde{I}_1, (\tilde{I}_1, u_1, u_2)) = q_1 = (1 - n_e p_{\text{end}})/(d_1^{**} L) \tag{4.5}$$

if the current realization has only one region so that a region add is required. We now need to change variables from $(\tilde{I}_1, u_1, u_2)$ to $\tilde{I}_2$. This is the dimension-matching step. The Jacobian for the change of variables is

$$\left|\frac{\partial(\ell_{1^*}, \ell_{2^*})}{\partial(u_1, u_2)}\right| = \begin{vmatrix} \frac{\partial \ell_{1^*}}{\partial u_1} & \frac{\partial \ell_{1^*}}{\partial u_2} \\ \frac{\partial \ell_{2^*}}{\partial u_1} & \frac{\partial \ell_{2^*}}{\partial u_2} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2$$

since $\ell_{1^*} = u_1 - u_2$ and $\ell_{2^*} = u_1 + u_2$. Hence

$$q(\tilde{I}_1, \tilde{I}_2) = 2\, q(\tilde{I}_1, (\tilde{I}_1, u_1, u_2))$$

for adding a mid region, with $q(\tilde{I}_1, (\tilde{I}_1, u_1, u_2))$ defined in equation 4.4 if the current realization has more than one region, or equation 4.5 if the current realization has more than one region.

If an end region add is being proposed, let ℓ_{1^*} be the location of the added switch in IBD status. If d_1^{**} is the maximum length of the proposed added region, sample u_1 uniformly from $(0, d_1^{**})$ and let ℓ_{1^*} equal u_1 if region is being added to the start of the chromosome, or $L - u_1$ if the region is being added to the end of the chromosome. The probability of proposing to add this particular end region is p_{end} if the current realization has only one region, or $\frac{1}{2}(1 - p_{\text{move}})p_{\text{end}}$ if the current realization has more than one region. The proposed realization $\tilde{I}_2$ is simply $(\tilde{I}_1, \ell_{1^*})$, so the probability density of the proposal is

$$q(\tilde{I}_1, \tilde{I}_2) = p_{\text{end}}/d_1^{**}$$

if the current realization has only one region, or

$$q(\tilde{I}_1, \tilde{I}_2) = \frac{1}{2}(1 - p_{\text{move}})p_{\text{end}}/d_1^{**}$$

if the current realization has more than one region.

The reverse of a proposal to add a region is a proposal to remove the added region. Let n_2^* be the number of regions on the *proposed* realization for which removal is permitted by the IBS data. The reverse probability density is

$$q(\tilde{I}_2, \tilde{I}_1) = \frac{1}{2}(1 - p_{\text{move}})/n_2^*.$$

When a region removal is proposed, the q-ratio is the multiplicative inverse of the q-ratio for the corresponding region add.

Metropolis-Hastings algorithm

At each step of the algorithm, a new continuous IBD process realization is proposed following the probabilities given above. The proposed process is accepted with probability a from equation 4.3. With probability $1 - a$ the proposal is rejected and the process remains in its current state.

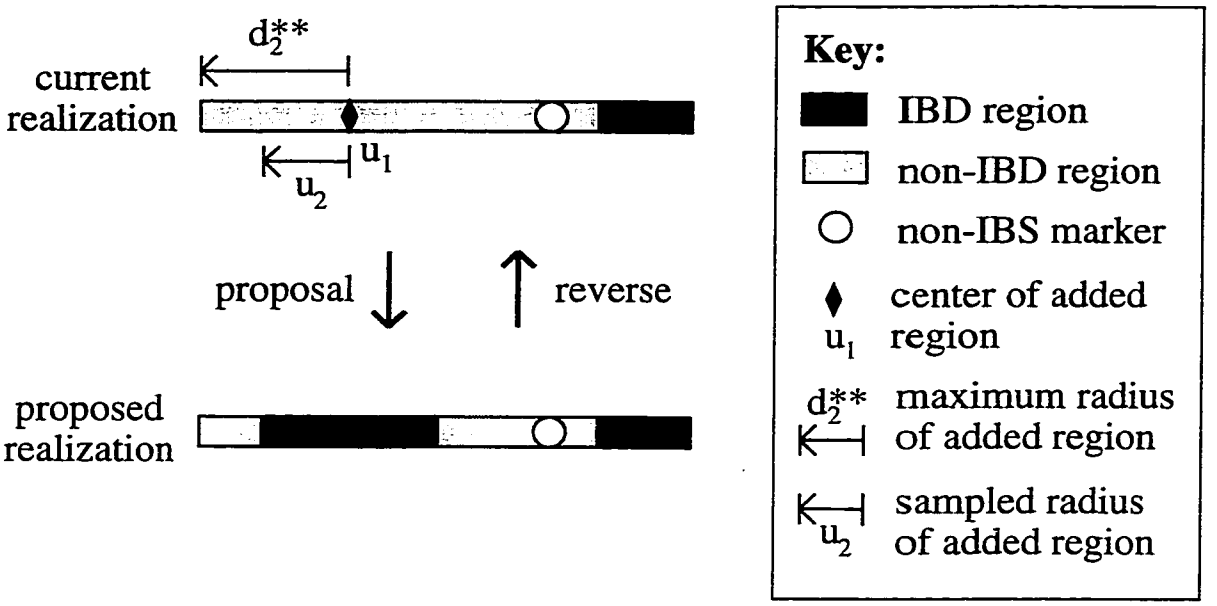

Figure 4.8: **Proposal to add a mid region, and reverse move.** The proposal is to add a region with center at u_1. The maximum radius of the added region is d_2^{**}. The sampled radius of the added region is u_2. The reverse move is removal of the second region of the proposed realization.

After each step of the algorithm, a Monte Carlo estimate $\hat{P}(\tilde{I})$ of the likelihood of the current continuous IBD process is found and the term $\hat{P}(\tilde{I})/P_A(\tilde{I})$ is calculated for the estimate given by equation 4.2.

4.2.4 Accounting for data errors

Most real IBS data will contain some genotyping errors which should be accounted for. If an error rate is known (or assumed), the possibility of errors can be incorporated into the likelihood calculation. These probabilities come into the calculation of P(D—I). Allowing for errors, the probability of non-IBS at a marker given underlying IBD has a small positive probability (rather than zero probability without allowing for errors), and the probability of IBS given IBD decreases slightly compared to the probability without allowing for errors. To allow for errors, we should allow IBD regions to cover non-IBS markers occasionally in the MCMC sampling procedure.

4.2.5 Use of common IBD process realizations

In section 4.2.2 it was noted that the normalizing constant cancels when calculating a likelihood ratio using the same choice of P_A in both numerator and denominator. In this case the sampling distribution (proportional to $P(\mathbf{D}|I)P_A(I)$) is the same for calculation of estimates of numerator and denominator of the likelihood ratio. One may either use the same sequence or two independent sequences of continuous IBD realizations in the estimates of numerator and denominator. Using one common sequence will save some computing time, and may increase or decrease the variance of the estimate of the ratio (see section 4.4 in Hammersley and Handscomb [18]). Let the estimates of numerator and denominator be X and Y with means μ_X and μ_Y. The variance of X/Y is approximately (by a Taylor series expansion, for example [20])

$$(\mathrm{Var}(X) - 2\mathrm{Cov}(X,Y)\frac{\mu_X}{\mu_Y} + \mathrm{Var}(Y)\frac{\mu_X^2}{\mu_Y^2})/\mu_Y^2.$$

Thus if the covariance of the two estimates is positive, using a common sequence will decrease the variance of the estimate of the ratio, while the variance will increase if the covariance is negative.

The covariance will be positive if the probability of an IBD realization under the model (model includes the crossing-over model and the relationship) for the numerator tends to be very close to the probability of the realization under the model for the denominator of the likelihood ratio. Generally, the probabilities will not be close if the relationships differ, but may be close if only the crossing-over model differs.

We present an example in section 4.4.1 for which it is best to use two independent sequences when calculating a likelihood ratio of aunt to half-aunt. In section 5.4.1 we find that a common sequence is best when calculating a likelihood ratio of chi-square ($M = 2$) model to Haldane's model with the half-cousins once removed relationship for the data considered there.

4.3 Bilateral relationships

Up to this point, we have considered identity by descent to be a binary process — two individuals are either identical by descent at a genetic locus, or they are not. However reality is a little more complex, because each of the two individuals has two copies of each gene and so, depending on the relationship, they may share two copies identical by descent, or one copy from one individual may be IBD with two copies of the other individual. Relationships in which only one parent of each individual is related to a parent of the other individual are called *unilateral*, and many of the relationships that we tend to consider, including half-sibs (half-brothers and half-sisters), cousins, and aunt-niece are indeed unilateral. *Bilateral* relationships are relationships in which the individuals are related through two distinct lines. Figure 4.9 shows two examples of bilateral relationships: sisters and double half-cousins.

We distinguish between bilateral relationships, in which the individuals are related

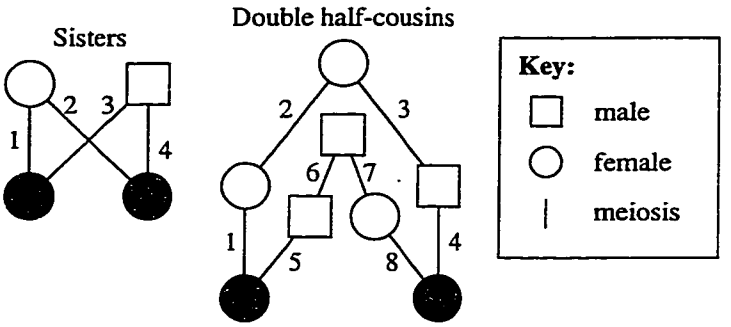

Figure 4.9: Examples of bilateral relationships.

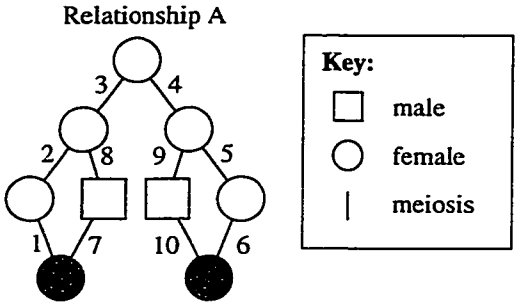

Figure 4.10: An inbred relationship.

by two lines that don't share any meioses, and other types of non-unilateral relationships, including inbred relationships in which the mother and father of at least one of the individuals are related to each other. For example, the sisters relationship in figure 4.9 has one line with meioses 1 and 2, and a second line with meioses 3 and 4, and the double half-cousins relationship has lines (1, 2, 3, 4) and (5, 6, 7, 8). Relationship A in figure 4.10 is inbred and lines (1, 2, 3, 4, 5, 6) and (7, 8, 3, 4, 9, 10) share meioses 3 and 4.

The methods described up to this point assume that we cannot distinguish between the cases of one or two copies IBD, which yields valid results but ignores some of the information in the data when the relationship under consideration is not unilateral. There is no simple way to deal with bilateral relationships directly in the Monte Carlo method for continuous IBD data described in chapter 2. However we can easily work

with bilateral relationships in the MCMC method for discrete IBS data described in this chapter.

Bilateral relationships can be incorporated into the MCMC procedure as follows. A bilateral relationship is made up of two unilateral relationships (these are the two lines) — for example the sibs relationship is made up of two half-sib relationships. Consider two continuous IBD processes $\tilde{I}^1$ and $\tilde{I}^2$, one for each line. In place of equation 4.1 write

$$\begin{aligned} P(\mathbf{D}) &= \int\int P(\mathbf{D}|\tilde{I}^1,\tilde{I}^2)P(\tilde{I}^1,\tilde{I}^2)d\tilde{I}^1 d\tilde{I}^2 \\ &= \int\int P(\mathbf{D}|\tilde{I})P_{A_1}(\tilde{I}^1)P_{A_2}(\tilde{I}^2)\frac{P(\tilde{I}^1,\tilde{I}^2)}{P_{A_1}(\tilde{I}^1)P_{A_2}(\tilde{I}^2)}d\tilde{I}^1 d\tilde{I}^2 \\ &\propto \int\int \frac{P(\tilde{I}^1,\tilde{I}^2)}{P_{A_1}(\tilde{I}^1)P_{A_2}(\tilde{I}^2)} \frac{P(\mathbf{D}|\tilde{I}^1,\tilde{I}^2)P_{A_1}(\tilde{I}^1)P_{A_2}(\tilde{I}^2)}{\int\int P(\mathbf{D}|\tilde{J}^1,\tilde{J}^2)P_{A_1}(\tilde{J}^1)P_{A_2}(\tilde{J}^2)d\tilde{J}^1 d\tilde{J}^2} d\tilde{I}^1 d\tilde{I}^2 \end{aligned}$$

The approximate probabilities P_{A_1} and P_{A_2} should correspond to the two unilateral relationships if possible (see section 4.2.1). Since the two lines don't share any meioses, the two IBD processes are independent, and $P(\tilde{I}^1,\tilde{I}^2) = P_1(\tilde{I}^1)P_2(\tilde{I}^2)$ where P_1 and P_2 are the probabilities of the IBD realizations given unilateral relationships 1 and 2 respectively. These probabilities can be estimated separately using the Monte Carlo likelihood of equation 2.3 as in section 4.2. Sampling of realizations of $\tilde{I}^1$ and $\tilde{I}^2$ should take place concurrently — at each step of the Metropolis Hastings procedure, propose a change to one of the IBD processes, chosen from the two processes at random. In choosing whether to accept the change, the probability $P(\mathbf{D}|\tilde{I}^1,\tilde{I}^2)$ depends on the joint IBD status of $(\tilde{I}^1,\tilde{I}^2)$ (i.e. zero, one or two copies IBD) following the formulas for probabilities of observed IBS data given the joint IBD status given in table 1 in [24].

For the normalizing constant to cancel out when calculating a likelihood ratio, it is important to use the same values for P_{A_1} and P_{A_2} in the numerator as in the denominator. If one of the relationships in the ratio is bilateral and the other is unilateral, the unilateral relationship needs to be thought of as a bilateral relationship

with one of the lines being the 'unrelated' relationship.

4.4 Evaluation of the MCMC procedure

In this section we will assess the Markov chain Monte Carlo procedure described above. To be useful, the procedure must give a good estimate in reasonable time. Parameters of the procedure should be fine-tuned to optimize acceptance probabilities of the Metropolis-Hastings algorithm and to otherwise decrease the variability of the procedure.

We will look briefly at the effects of choosing different levels of the main parameters of the procedure, comparing the results against a baseline choice of parameters. The parameters that we will examine are: the choice of relationship used in the approximate probability calculations P_A; the values of p_{move} and p_{end} used in generating proposals; and the number of Monte Carlo iterations used in calculating each estimate $\hat{P}_i(\tilde{I}_i)$.

For this test, we will use marker data on chromosome 1 for an aunt niece pair, individuals 5 and 6 from family 120 of the Mennonite study described in chapter 5 (see figure 5.1). Chromosome 1 has 37 markers, and the distance from first marker to last is 2.78 Morgans. We will estimate the likelihood ratio of aunt to half-aunt under Haldane's model. The likelihood ratio, which we can calculate exactly using the extension to Boehnke and Cox's method described in section 2.4.1, equals 0.3314.

Table 4.1 records the average changes to the performance of the MCMC procedure for different values of the parameters of the procedure. The parameters are described relative to the following baseline: P_A based on the aunt relationship, $p_{\text{move}} = 0.5$, $p_{\text{end}} = 0.1$, 100,000 MCMC iterations, 1 Monte Carlo iteration used in calculating each estimate $\hat{P}_i(\tilde{I}_i)$, and independent sequences of IBD realizations used in estimating the numerator and denominator of the likelihood ratio.

Performance results for the baseline are given in the first row of the table. The per-

formance measures considered are: acceptance probability for the Metropolis-Hastings ratio, average estimated standard error, observed standard error, CPU time, and a time factor which is the observed variance (square of observed standard error) multiplied by CPU time. The standard error is estimated by the method of batch means [13], with batches of size 1000, and the average is over 100 separate runs. The observed standard error is the standard deviation of the likelihood ratio estimates from 100 runs. Acceptance probabilities are an average of the observed acceptance percentages of 100 runs. CPU time is an average from 5 runs on a DEC Alphastation 400-233 with 192 MB RAM.

We would like to see high acceptance probabilities, low observed standard errors and low CPU times. Estimated standard errors tend to be biased downward; we can assess the degree of bias by comparing the average estimated standard errors to the observed standard errors. To balance the requirements for low standard errors and low CPU times, we look at the time factor.

The time factor gives an impression of the relative times that the procedure will take to produce a given level of standard error. The variance of the estimator is inversely proportional to the number of MCMC iterations, and hence is also inversely proportional to CPU time (excepting set-up time and time for burn-in MCMC iterations), if all other parameters are fixed. For example, for the baseline parameters, the time factor is 0.22. Hence, to achieve a standard error of 0.001 (a variance of 10^{-6}), the number of MCMC iterations would have to be increased to a level that would cause the procedure to take about $0.22/10^{-6} = 2.2 * 10^5$ seconds, or 61 hours. On the other hand, with the change to 100 MC iterations per estimate of $\hat{P}_i(\tilde{I}_i)$, the time factor is 0.0056, so to achieve a standard error of 0.001, the number of MCMC iterations would have to be increased to a level that would cause the procedure to take about $0.0056/10^{-6} = 5.6 * 10^3$ seconds, or 1.6 hours. The ratio of the time factors ($0.0056/0.22 = 0.03$) gives the relative amounts of time that the procedure will take to achieve the same standard error (e.g. 1.6 hours/61 hours = 0.03) with these choices

Table 4.1: **Performance measures for the MCMC procedure.** Performance measures and baseline are described in the text. Line 1 gives results for the baseline. Lines 2 and 3 vary $p_{\mathbf{move}}$ from the baseline value of 0.5. Lines 4 and 5 vary $p_{\mathbf{end}}$ from the baseline 0.1. Lines 6, 7 and 8 vary the number of Monte Carlo iterations used in calculating the estimate $\hat{P}_i(\tilde{I}_i)$ (the baseline is 1 Monte Carlo iteration). Lines 9 and 10 use different choices of relationship in the approximation P_A. The baseline uses the aunt relationship, line 9 uses the half-aunt relationship, and line 10 uses a mixture of the half-aunt and aunt relationships.

	change from baseline	acceptance probability	ave. est. std. err.	observed std. err.	CPU time (seconds)	time factor
1	**no change**	43%	0.064	0.062	57	0.22
2	$\mathbf{p_{move} = 0.4}$	43%	0.062	0.063	56	0.22
3	$\mathbf{p_{move} = 0.6}$	43%	0.062	0.074	59	0.23
4	$\mathbf{p_{end} = 0.05}$	41%	0.062	0.063	55	0.22
5	$\mathbf{p_{end} = 0.2}$	44%	0.066	0.074	59	0.32
6	**10 MC**	43%	0.022	0.020	59	0.024
7	**100 MC**	43%	0.0085	0.0081	85	0.0056
8	**1000 MC**	43%	0.0058	0.0058	334	0.011
9	**half-aunt P_A**	35%	0.069	0.084	39	0.28
10	**mixture P_A**	38%	0.066	0.066	45	0.20

of parameters, with the appropriate increases in the number of MCMC iterations.

The procedure will be most efficient when the time factor is minimized. From table 4.1 it is clear that the most critical parameter is the number of MC iterations per estimate of $\hat{P}_i(\tilde{I}_i)$. Let $\tilde{n}$ be the number of MC iterations per estimate of $\hat{P}_i(\tilde{I}_i)$. We expect CPU time to increase linearly with the number of MC iterations, and the least squares line

$$\text{CPU time} = 56.7 + 0.2773\,\tilde{n}$$

fits the time data very well. We also expect the part of the variance of the estimates of the likelihood ratio that is attributable to the variability in the estimates $\hat{P}_i(\tilde{I}_i)$ to decrease as $1/\tilde{n}$. Thus we expect a relationship of the form $\text{var} = a + b/\tilde{n}$, where a is the part of the variance that is attributable to the variability in the MCMC sampling of continuous IBD processes. Fitting the line

$$\tilde{n}\text{var} = a\tilde{n} + b$$

by least squares gives $a = 2.993 * 10^{-5}$ and $b = 3.699 * 10^{-3}$. Hence we estimate the relationship between the time factor and the number of MC iterations to be

$$\text{var} * \text{CPU time} = (2.993 * 10^{-5} + 3.699 * 10^{-3}/\tilde{n})(56.7 + 0.2773\,\tilde{n}). \qquad (4.6)$$

The expression is minimized by setting the number of MC iterations to 159, and the estimated minimum time factor is 0.0054. As a check, the procedure was performed with 160 MC iterations, which gave an observed standard error of 0.0070 and CPU time of 105, so that the time factor is 0.0051, which is slightly less than the estimate.

Figure 4.11 plots the estimated relationship between the time factor and the number of MC iterations. It is apparent from the plot that while the estimated minimum time factor is achieved at 159 MC iterations, the curve is quite flat around that number. Any number of MC iterations between 81 and 315 will give an estimated time factor less than 0.006 which is not much different from the estimated minimum time factor of 0.0054.

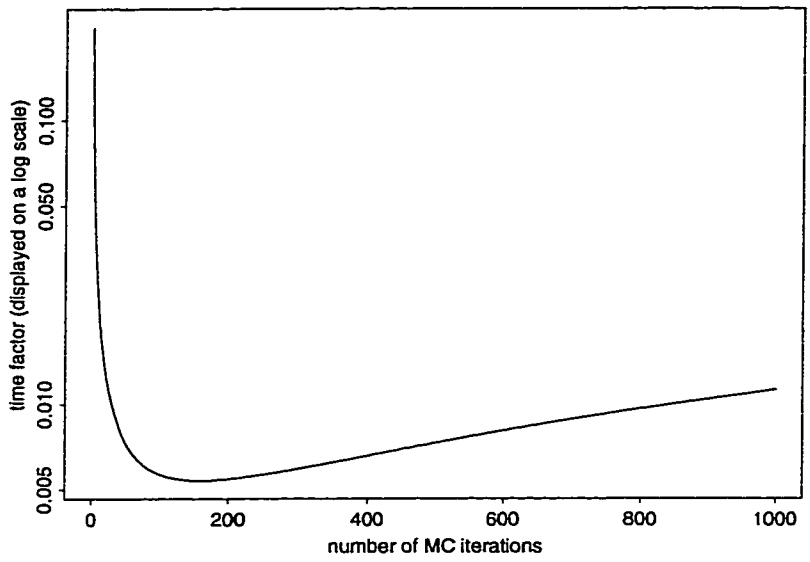

Figure 4.11: Estimated time factor curve.

4.4.1 Evaluation of the use of common IBD process realizations

When calculating the likelihood ratio of aunt to half-aunt, using the data and baseline parameters described above, the standard error is approximately 0.062 when using two independent sequences and approximately 0.070 when using one common sequence, while the computation time is 57 seconds for two sequences and 50 for one sequence. The time factor is thus 0.22 for two sequences but 0.25 for one sequence, hence it is better to use two independent sequences in this case.

Chapter 5

DATA ANALYSIS

In this chapter we apply the MCMC method described in chapter 4, as well as the extension to Boehnke and Cox's [3] method described in section 2.4.1 to analysis of a real data set. The data (Erik Puffenberger and Aravinda Chakravarti, unpublished data) consist of microsatellite marker genotype information on 100 individuals organized into small families, mostly trios consisting of two parents with one child, as shown in figure 5.1.

On the 22 autosomal chromosomes there are a total of 589 markers in the data set. To assign locations to the markers we used the map developed by Karl Broman in 1998. This map can be found at *http://www.marshmed.org/genetics/* and is described in [5]. Of the 589 markers, 32 markers were not found on Broman's map and were discarded. Of any group of markers attributed to the same location by Broman's map, only the first of those markers was used, resulting in a further 39 markers being discarded. Thus in total 488 markers were used, which averages to a little more than 10 markers per Morgan. Throughout this analysis we assume linkage equilibrium and use as population allele frequencies the observed allele frequencies from the data.

Since the number of individuals is large, it is not possible to fully analyze all 4950 possible pairs of individuals. Hence the analysis is developed in several stages with each stage analyzing fewer pairs but in greater depth.

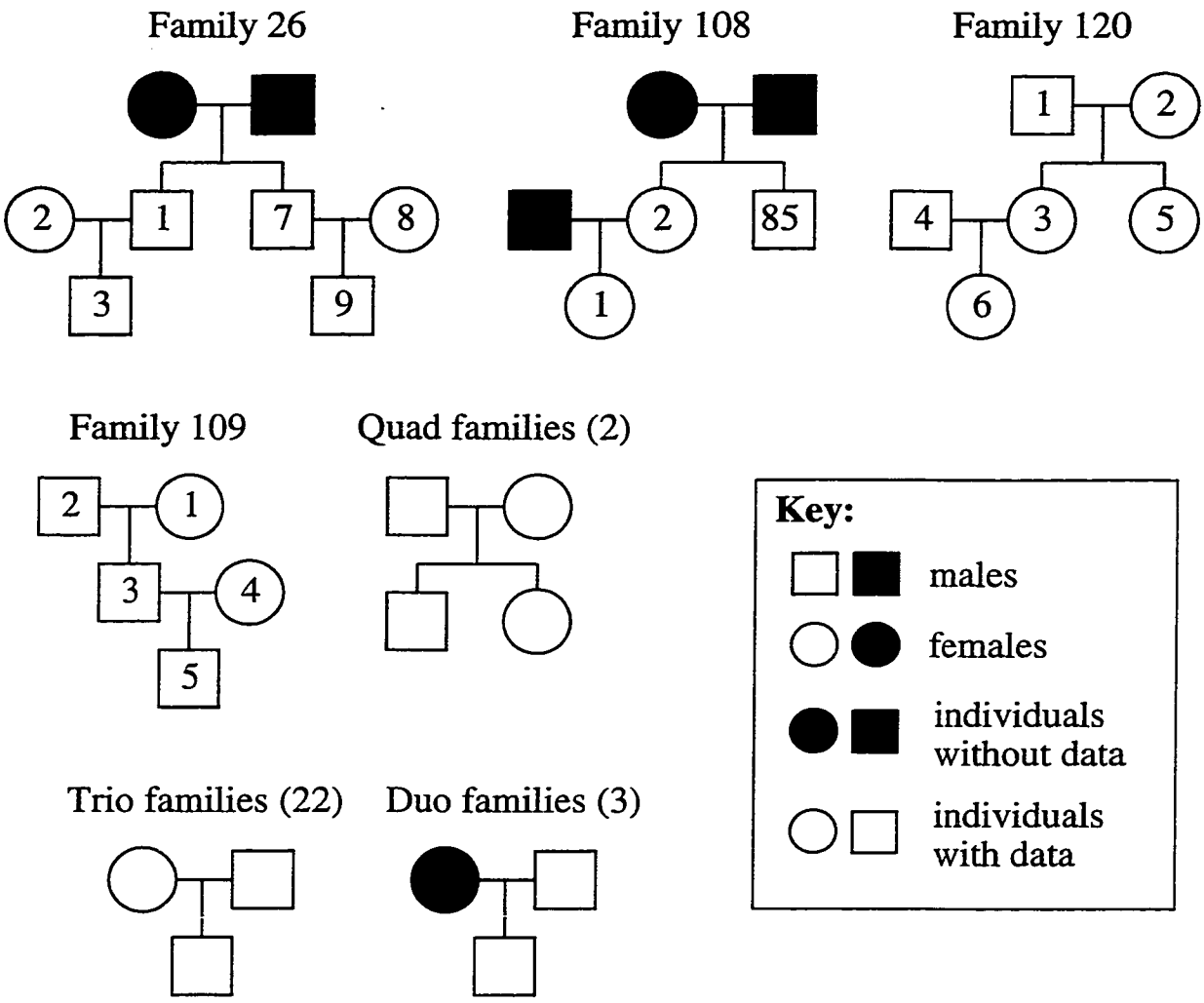

Figure 5.1: Mennonite families.

5.1 Preliminary analysis

The individuals used for the first stage of the analysis consist of all individuals except those for whom there were data on both parents. For example, the children of the trio families were not included in the analysis. Of those individuals included in the analysis, all possible pairs of individuals were considered, except for pairs in which both individuals belong to the same family. Hence the total number of pairs considered was 2169.

Five 'half-sib-type' (see section 2.4.2) relationships were considered first — half-sibs, half-aunt, half-cousins, half-cousins once removed and half second cousins. The likelihood ratios of each of the relationships versus unrelated were calculated assuming Haldane's model and using the extension to Boehnke and Cox's method (section 2.4.1). The computation time of the method increases exponentially with the number of meioses, and, since we wanted to consider some 'distant' relationships with low expected proportion of IBD, we chose half-sib-type relationships because for a given number of meioses they minimize the expected proportion of IBD.

Statistical geneticists tend to use LOD scores (base ten log of the likelihood ratio) to assess statistical significance, and usually use 3.0 as a cut-off for presumed significance. Twenty pairs of individuals had LOD scores greater than 3.0 for at least one of the relationships considered. We will assess the significance of these results in section 5.3.

The twenty pairs of individuals can be grouped into nine 'discovered families' (if A seems to be related to B and B to C then A, B and C make up one discovered family). Of the relationships considered, the relationship with the highest LOD score for each of these twenty pairs was half-sibs for five pairs, half-aunt for eight pairs, half-cousins for six pairs and half-cousins once removed for one pair.

These results suggest that the relationships with the highest LOD scores tend to be the close relationships. To find the more distant relationships it may be necessary

to examine pairs with LOD scores lower than 3.0.

5.2 *Further analysis of twenty pairs of individuals*

We now take a closer look at the twenty pairs singled out in the preliminary analysis, widening the pool of possible relationships to include siblings (brothers and sisters), aunt-niece, great-aunt, grandparent, great grandparent, cousins, cousins once removed, and second cousins. LOD scores for each of these relationships versus unrelated were calculated using the extension to Boehnke and Cox's method with orbits (see section 2.4.1) and assuming Haldane's model.

Table 5.1 presents the results, showing the two highest LOD scores for each pair of individuals. Individuals are identified by a family and individual identification number, for example individual 40-1 is individual 1 from family 40. Horizontal lines separate the 'discovered families'. Note that individuals with the same family identification number (such as individuals 40-1 and 40-7) need not be related to each other, as individuals that have married into a family are considered family members.

The LOD scores are the base ten log of the ratio of the likelihood of the relationship to the likelihood of unrelated. Relationships corresponding to the highest LOD scores are indicated by abbreviations which are explained in table 5.2. For example, the most likely relationship between the individuals 40-1 and 51-2 is siblings, with LOD score 22.41, and the second most likely relationship is grandparent, with LOD score 19.12.

Table 5.1 also shows the amount of missing data for each pair of individuals — if one individual of the pair does not have data at a marker, then the other individual's data at that marker do not provide any information about the relationship and the marker is scored as missing data. It is interesting that some pairs of individuals have large amounts of missing data yet very high LOD scores (for example, 368 markers out of 488 are missing for pair 40-1 51-2 while the LOD score for siblings versus unrelated is 22.4) showing that the remaining markers provide a lot of information about the

relationship.

In addition, table 5.1 shows the percentage of IBS markers for the pairs of individuals. A marker is IBS if one or more alleles in one individual are shared by the other individual. Given the allele frequencies in the data set, a pair of unrelated individuals will be IBS at 63% of their markers on average. High IBS percentages indicate relatedness, but it is evident from the table that IBS percentage alone does not provide much information about relationship compared with the information provided by the full likelihood (represented by the LOD scores). For example the pair of individuals 108-85 and 113-2 have a LOD score of 4.16 for the half-cousin relationship (which, as we will see in section 5.3, shows that there is sufficient evidence to reject the possibility that they could be unrelated), while their IBS percentage is only 65%, barely above the average for unrelated individuals.

5.2.1 Issues pertaining to relationship inference

Even if the human genome were infinitely long, IBS or IBD data alone would not be sufficient to distinguish between all relationship possibilities. There are classes of relationships that are indistinguishable on the basis of their IBD processes (and hence IBS data) because the relationships have the same number of meioses and the meioses can be labeled in such a way that the IBD states are the same. For example, half second cousins are indistinguishable from half-cousins twice removed and from half-sibs four times removed, as in figure 5.2. Also, using the genotype data we cannot tell which individual is which member of the relationship — for example if two individuals are half-aunt and half-niece we can't use the data to tell us which individual is the aunt and which is the niece. These issues could mostly be resolved by looking at the ages of individuals.

We have not taken sex into account when analyzing relationship. For example, if two males are recorded as having highest LOD score for aunt-niece it should be taken that in fact their highest LOD is for uncle-nephew.

Table 5.1: Analysis of twenty pairs of individuals.

Pair of individuals		Highest LOD		Second highest		Missing Data /488	IBS Percentage
40-1	51-2	22.41	s	19.12	gp	368	93%
40-1	110-2	34.36	a	33.75	hs	55	85%
51-2	110-2	15.15	a	14.85	hs	364	86%
40-7	93-1	5.14	ha	5.10	ggp	40	69%
40-7	94-1	5.32	cr	5.26	hc	39	69%
49-2	94-1	6.20	ga	6.19	c	32	70%
50-2	52-1	91.13	s	72.87	hs	24	89%
51-1	113-2	3.99	cr	3.91	hc	22	71%
108-85	113-2	4.16	hc	4.02	cr	31	65%
52-2	113-1	17.59	c	17.47	a	71	76%
54-2	109-1	4.32	hc	4.23	cr	36	66%
93-2	98-2	15.10	a	15.02	hs	199	84%
93-2	109-1	5.66	c	5.11	ga	33	69%
97-1	109-4	3.06	hcr	3.04	sc	43	69%
107-2	112-2	15.59	c	15.56	ga	102	78%
107-2	115-1	3.04	hc	2.94	hcr	41	68%
115-1	124-1	10.60	ggp	10.48	ha	64	74%
115-1	124-2	8.46	c	8.41	ha	47	69%
124-2	125-1	4.68	c	4.42	ga	72	73%
120-4	122-1	4.49	cr	4.40	hc	68	67%

Table 5.2: Abbreviations for relationship names.

Relationship	**Abbreviation**
half-sibs	hs
half-aunt	ha
half-cousins	hc
half-cousins once removed	hcr
half second cousins	hsc
siblings	s
aunt	a
great-aunt	ga
cousins	c
cousins once removed	cr
second cousins	sc
grandparent	gp
great grandparent	ggp

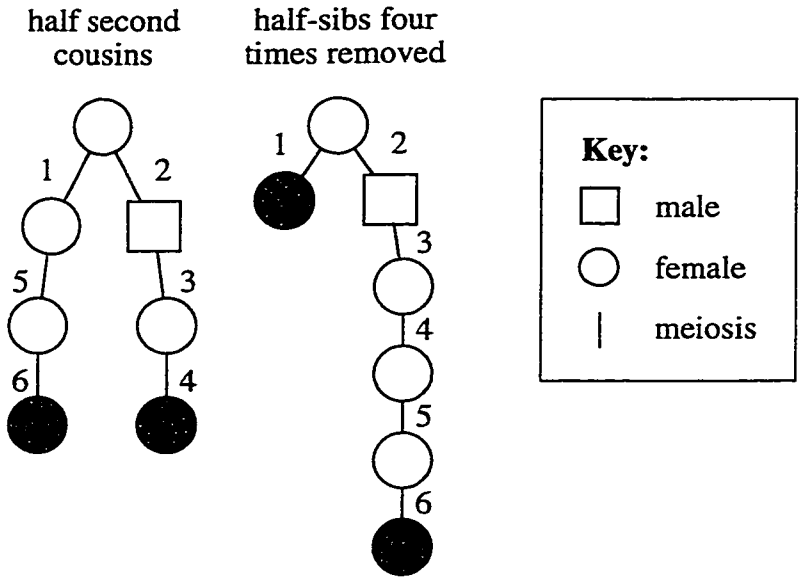

Figure 5.2: Indistinguishable relationships.

In this analysis, each pair of individuals was analyzed independently of other pairs. In section 5.3 we will see that it is often not possible to distinguish between the relationships with the two highest LOD scores for a pair of individuals. Hence the relationships with the highest LOD scores need not be consistent within a discovered family. For example, if A is a sister (sibling) of B and B is a sister of C then A must be a sister of C, however if the highest LOD scores for the relationships between A and B and between B and C are for siblings, the highest LOD score for the relationship between A and C need not be for siblings. However, it should also be noted that if A is related to B and B to C it is not always true that A is related to C. For example, A may be related to B through B's mother, while B may be related to C through B's father, in which case A is not related to C.

5.3 Assessing significance and goodness-of-fit

In assessing significance of the LOD scores for a given relationship versus unrelated we will simulate data assuming, as in the above analysis, that the marker loci are all in linkage equilibrium, so that the allele probabilities are independent from locus to locus, and that the observed allele frequencies in the data are the population frequencies. These assumptions are unrealistic but are necessary since we do not have data on additional pairs of unrelated individuals.

Let us consider a hypothesis test in which the null hypothesis is that a given pair of individuals is unrelated, and the alternative hypothesis is that the pair of individuals is related with relationship R (R may be siblings, aunt-niece, etc.). In order to evaluate the hypotheses, we calculate LOD(R vs. unrelated) from IBS data on the individuals, and reject the null hypothesis if the LOD score exceeds some critical value. To determine the critical value, suppose we had IBS data on an infinite number of unrelated individuals and could calculate LOD(R vs. unrelated) for each one. The critical value for a level α test would be the $(1-\alpha)100\%$ quantile of the LOD scores on the unrelated individuals. To estimate this quantile, we simulate IBS data for a large number of pairs of unrelated individuals, and take the $(1-\alpha)100\%$ sample quantile.

We estimated the 80% and 95% quantiles from LOD scores calculated on one thousand simulated sets of IBS data from unrelated individuals. The simulated data did not have any markers with missing data. Columns 2 and 3 on table 5.3 show the estimated quantiles. For example, the estimated 95% quantile of LOD(half-sibs vs. unrelated) for unrelated individuals is -11.7, and hence the estimated critical value for rejecting the hypothesis of unrelated in a test of unrelated vs. half-sibs at a 5% significance level is -11.7. In practice, we would not reject the unrelated hypothesis unless the LOD score was greater than zero, since a LOD score less than zero means that it is more likely that the individuals are unrelated than that they are related

with relationship R. Thus an estimated $(1-\alpha)100\%$ quantile less than zero effectively means a critical value of zero for the level α test.

The quantiles tend to be higher for distant relationships (e.g. half second cousins) than for close relationships (e.g. siblings), so that a higher LOD score is required for statistical significance for distant relationships. The highest 95% quantile is 0.4 for half second cousins, so that with no missing data, a critical value much less than 3 can be used to assess significance. In particular, all of the LOD scores in table 5.1 are significant at the 5% level. For many of the relationships, the estimated 95% quantile is less than zero, so that if the LOD score is greater than zero we can reject the hypothesis that the individuals are unrelated at a significance level of 5%.

5.3.1 Significance with missing data

In the missing data column of table 5.1, we see that several of the twenty pairs of individuals have large amounts of missing data (up to 368 markers with missing data out of 488 markers in total). Missing data reduce the amount of information about the relationship in the data, and critical values for a test of unrelated versus related with relationship R will be higher with missing data than without. Equivalently, 80% and 95% quantiles of the LOD(R vs. unrelated) scores for unrelated pairs will be higher (in most cases, this means less negative) with missing data than with no missing data.

To investigate the effect of missing data on the quantiles of the LOD scores for unrelated individuals, we simulated pairs of unrelated individuals with the same missing data patterns as pairs of analyzed individuals and estimated LOD score quantiles. For each missing data pattern, we did not estimate quantiles of LOD(R vs. unrelated) for all relationships R, but only for the two relationships with highest LOD scores for the pair from whom the missing data pattern came. For example, the two highest LOD scores for pair 40-1 51-2 are for siblings and grandparent relationships (from table 5.1) so for the missing data pattern corresponding to this pair we only estimated quantiles

Table 5.3: Sample quantiles of LOD(R vs. unrelated|Haldane's model) for pairs of simulated unrelated and related individuals with no missing data. The unrelated sample quantiles are used as critical values for a test of unrelated vs. related with relationship R, while the related sample quantiles are used to assess power and goodness of fit.

	unrelated		**related (R)**	
Relationship (R)	**80%**	**95%**	**5%**	**50%**
half-sibs	-13.5	-11.7	27.6	42.4
half-aunt	-3.9	-2.9	5.0	14.2
half-cousins	-1.2	-0.4	-0.0	4.4
half-cousins once removed	-0.3	0.3	-0.7	1.2
half second cousins	0.0	0.4	-0.5	0.2
grandparent	-10.5	-9.1	29.3	47.7
great grandparent	-3.7	-2.7	6.0	16.3
siblings	-33.9	-31.7	111.6	152.0
aunt	-14.4	-12.6	26.8	40.9
great-aunt	-4.0	-2.9	4.8	13.8
cousins	-4.3	-3.1	5.0	13.1
cousins once removed	-1.2	-0.3	0.0	3.9
second cousins	-0.3	0.3	-0.7	1.2

for LOD(siblings vs. unrelated) and LOD(grandparent vs. unrelated). These results can be used to assess the statistical significance of the LOD scores in table 5.1.

Of the twenty pairs, pair 40-1 51-2 has the most markers with missing data (368). We analyzed one thousand pairs of simulated unrelated individuals with missing data at the same 368 markers. The missing data estimated LOD quantiles for siblings are -11.6 (80%) and -9.6 (95%), compared to -33.9 and -31.7 with no missing data, and for grandparent the missing data estimated quantiles are -3.6 (80%) and -2.3 (95%), compared to -10.5 and -9.1 with no missing data. The estimated 80% and 95% quantiles for these relationships are quite a bit higher with missing data than without, but are still less than zero, so the effective critical values (at 5% and 20% significance levels) remain at zero. The tests siblings vs. unrelated and grandparent vs. unrelated both reject the possibility of unrelated at the 5% level since the LOD scores LOD(siblings vs. unrelated)=22.41 and LOD(grandparent vs. unrelated)=19.12 (from table 5.1) for this pair are both greater than zero.

Pair 51-1 113-2 has the fewest missing markers (22) of the twenty pairs. We analyzed one thousand pairs of simulated unrelated individuals with missing data at the same 22 markers. The missing data estimated LOD quantiles for cousins once removed are -1.1 (80%) and -0.3 (95%), compared to -1.2 and -0.3 with no missing data, and for half-cousins the missing data estimated quantiles are -1.2 (80%) and -0.4 (95%), compared to -1.2 and -0.4 with no missing data. Thus this small amount of missing data has essentially no effect on the quantiles. Tests of cousins once removed vs. unrelated and half-cousins vs. unrelated both reject the possibility of unrelated at the 5% level

Pair 40-1 110-2 has a moderate amount of missing data (55 markers). Again we analyzed one thousand pairs of simulated unrelated individuals with missing data at these 55 markers. The missing data estimated LOD quantiles for aunt-niece are -13.6 (80%) and -11.3 (95%), compared to -14.4 and -12.6 with no missing data, and for half-sibs the missing data estimated quantiles are -12.8 (80%) and -10.7 (95%),

compared to -13.5 and -11.7 with no missing data. This amount of missing data has had some effect on the quantiles, but the critical values for these relationships at 5% and 20% significance levels remain at zero. Again, tests of aunt vs. unrelated and half-sibs vs. unrelated both reject the possibility of unrelated at the 5% level.

The quantiles presented above are estimates and are subject to sampling variability. To estimate this sampling variability, we performed a bootstrap procedure on the sampled LOD scores. From a set of LOD scores from one thousand simulated individuals, 1000 were sampled with replacement, for 1000 repetitions. The estimated $(1-\alpha)100\%$ quantiles from each repetition were sorted and the 50th and 950th of the 1000 sampled values form a 90% bootstrap confidence interval for the $(1-\alpha)100\%$ quantile. For the 95% quantile of the LOD(half-sibs versus unrelated) scores for unrelated pairs, a 90% bootstrap confidence interval is (-11.9, -11.5). For the 80% quantile of the LOD(half-cousins vs. unrelated) scores for unrelated pairs, a 90% bootstrap confidence interval is (-0.5,-0.2). Thus the sampling error in the estimated quantiles may be up to approximately 0.2.

5.3.2 Goodness-of-fit and power

To assess the goodness-of-fit of the relationships with the highest LOD scores and to determine power to distinguish whether a pair of individuals is related or unrelated, we again use data simulated assuming linkage equilibrium and using observed allele frequencies from the data. One thousand pairs of individuals were simulated from each relationship, and LOD(R vs. unrelated) was calculated for the true relationship R for each pair.

Columns 4 and 5 of table 5.3 show estimated 5% and 50% quantiles of the LOD(R vs. unrelated) scores when R is the true relationship. For close relationships (siblings, half-sibs, etc) the estimated 5% quantile when R is true is much greater than the estimated 95% quantile when the individuals are unrelated. Hence for these close relationships, it will almost always be possible to tell whether a pair of individuals

are unrelated or related according to one of these close relationships. However for more distant relationships, such as half second cousins and half-cousins removed, it will not always be possible to tell whether a pair of individuals is unrelated or has one of these distant relationships. In particular, the estimated 95% quantile for LOD(half second cousins versus unrelated) is 0.4 for unrelated individuals, while the estimated 50% quantile for the LOD scores is 0.2 when the individuals are half second cousins, so that at a 5% significance level it will not be possible to tell that half second cousins are not unrelated in more than half of such cases. For comparison, Donnelly [7] shows that with continuous IBD data on all 22 autosomal chromosomes it is possible to tell that half second cousins (equivalently half-sibs four times removed) are not unrelated in approximately 99.7% of such cases (see table I in [7]) assuming Haldane's model and a total map length of 33 Morgans.

We can check that the highest LOD scores in table 5.1 are consistent with the distribution of LOD scores that we expect to see if the relationship is true. For example, for individuals whose true relationship is aunt-niece, we expect to see a LOD(aunt vs. unrelated) score higher than 26.8 in 95% of such cases. In table 5.1, pair 40-1 110-2 has a LOD score of 34.36 for aunt, which is reasonable, but pairs 51-2 110-2 and 93-2 98-2 have LOD scores lower than 26.8 for aunt. However these two pairs have large amounts of missing data, which reduces the expected LOD scores.

5.3.3 Distinguishing between relationships

We can also investigate whether it is possible to distinguish between the relationships with the highest and second highest scores of LOD(R versus unrelated). As an example, we look at whether it is possible to distinguish between aunt and half-sibs. For 1000 pairs of individuals sampled from the aunt relationship we calculated LOD(aunt vs. half-sibs) which equals LOD(aunt vs. unrelated) minus LOD(half-sibs vs. unrelated), and we did the same for 1000 pairs of individuals sampled from the half-sibs relationship. The estimated 95% quantile of the LOD scores from the half-sibs data is

0.38. Thus a difference in LOD(aunt vs. unrelated) and LOD(half-sibs vs. unrelated) of at least 0.38 is required to reject the possibility of half-sibs in favor of aunt at the 5% significance level with no missing data — a greater difference will be required if some data are missing. The estimated 80% quantile of the LOD scores from the aunt data is 0.37, thus it will not be possible to distinguish between aunt and half-sibs at a 5% significance level when the data are from an aunt and niece in at least 80% of such cases. The difference in LOD scores is not significant for pairs 51-2 110-2 (difference 0.3) or pair 93-2 98-2 (difference 0.08). The difference in LOD scores is 0.61 for pair 40-1 110-2, which may be a significant difference, however an analysis incorporating the 55 markers with missing data would be required to accurately determine significance for this pair.

5.4 Beyond Haldane's model

To investigate the effect on the analysis of assuming a crossover model other than Haldane's, we reanalyzed several pairs of individuals calculating probabilities under the chi-square ($M = 2$) model.

Consider the LOD score for some relationship R versus unrelated. The likelihood of unrelated is not affected by the crossing-over model, so the LOD score when the chi-square model is assumed is

$$\begin{aligned}
&\text{LOD}(R \text{ vs. unrelated} \mid \text{chi-square model}) \\
&\quad = \log_{10}(L(R, \text{ chi-square model})) - \log_{10}(L(\text{unrelated})) \\
&\quad = \log_{10}(L(R, \text{ chi-square model})) - \log_{10}(L(R, \text{ Haldane's model})) \\
&\qquad + \log_{10}(L(R, \text{ Haldane's model})) - \log_{10}(L(\text{unrelated})) \\
&\quad = \text{LOD}(\text{chi-square model vs. Haldane's model} \mid R) \\
&\qquad + \text{LOD}(R \text{ vs. unrelated} \mid \text{Haldane's model})
\end{aligned}$$

Thus to calculate LOD(R vs. unrelated | chi-square model), we can calculate

LOD(chi-square model vs. Haldane's model $|\, R$) using the MCMC method of section 4.2 and add it to the LOD(R vs. unrelated $|$ Haldane's model) already calculated for the results in section 5.2.

The pairs of individuals and relationships analyzed with the chi-square model will be pair 40-1 110-2 for the aunt and half-sib relationships, pair 40-1 51-2 for the siblings and grandparent relationships, and pair 97-1 109-4 for the half-cousins once removed and second cousin relationships. The first and second pairs are closely related while the third pair is more distantly related. Both relationships for the first and third pairs have the same expected IBD proportion, while the two relationships for the second pair have different expected IBD proportions.

For all the analyses in this section we use $p_{\text{move}} = 0.5$, $p_{\text{end}} = 0.1$, and the relationship of interest for the approximate distribution P_A. The parameters p_{move} and p_{end} are introduced in section 4.2.3, and P_A is discussed in section 4.2.1. Choice of these parameters is investigated in section 4.4. Choice of remaining parameters for the calculations is discussed below in section 5.4.1. For each of the analysis we used one million MCMC iterations and 250 Monte Carlo iterations per estimate $\hat{P}_i(\tilde{I}_i)$. However for some analyses we used a common sequence of IBD realizations while for others two separate sequences were used. For the aunt and for the half-sib analysis of pair 40-1 110-2 we used two separate sequences of IBD realizations. We also used two separate sequences for the siblings and grandparent analyses of pair 40-1 51-2. For half-cousin once removed and second cousin analysis of the pair 97-1 109-4 we used one common sequence of IBD realizations.

Standard errors are estimated as follows. Let X and Y be the estimated numerator and denominator of the likelihood ratio, obtained from the MCMC procedure. The standard errors s_X and s_Y of X and Y are estimated using batch means from the MCMC run, with 1000 MCMC iterations per batch. If a common sequence of IBD realizations is used in estimating the numerator and denominator, the estimated covariance $s_{X,Y}$ between X and Y is also estimated by batch means. Then (see section 4.2.5),

the estimated standard error s_Z of the ratio $Z = X/Y$ is $\sqrt{s_X^2 + s_Y^2 Z^2 - 2s_{X,Y}Z}/Y$ where $s_{X,Y}$ is zero if separate sequences of IBD realizations were used in numerator and denominator. The estimated LOD score is $\log_{10}(Z)$, and the estimated standard error s_{LOD} of the LOD score is $s_Z/(\ln(10)Z)$ (from a Taylor series expansion of $\log_{10}(Z)$ about $E(Z)$, see for example [20]). In this manner, standard errors are estimated for the LOD scores estimated for each chromosome, and the final LOD, which is the sum of the LOD scores over all 22 chromosomes, has estimated standard error being the square root of the sum of all the s_{LOD} squared.

5.4.1 Choice of parameters for calculations

To investigate choice of parameters for the calculation procedure for the aunt relationship, we ran a preliminary analysis on the single chromosome of aunt data used in section 4.4. Table 5.4 shows the results. We estimate the likelihood ratio of chi-square ($M = 2$) model with aunt relationship versus Haldane's model with aunt relationship to be approximately 0.6. The observed standard error is the standard deviation of estimates of the likelihood ratio over 100 runs, and the CPU time is the average CPU time in seconds from 3 runs. As in section 4.4, the time factor equals the square of observed standard error multiplied by CPU time. Each run consisted of 100,000 iterations of the MCMC procedure using the aunt relationship in the approximate probability P_A, $p_{\text{move}} = 0.5$ and $p_{\text{end}} = 0.1$. The procedure was repeated with and without use of a common sequence of IBD realizations in numerator and denominator of the ratio (see section 4.2.5), and with 10 and 100 Monte Carlo iterations for each estimate of $\hat{P}_i(\tilde{I}_i)$ (see section 4.2).

From the results in table 5.4 we see that using a common sequence of IBD realizations yields a lower standard error than using two independent sequences when only 10 Monte Carlo iterations are used per estimate $\hat{P}_i(\tilde{I}_i)$. However when the number of Monte Carlo iterations per $\hat{P}_i(\tilde{I}_i)$ is increased to 100, using two sequences gives a lower standard error and overall results in a lower time factor than using a common

Table 5.4: Investigation of choice of parameters for crossover model analysis for aunt.

	common sequence		two sequences	
	10 MC	100 MC	10 MC	100 MC
observed s.e.	0.18	0.095	0.19	0.068
CPU time	55	83	62	91
time factor	1.8	0.75	2.3	0.42

sequence. Using these results to estimate the time factor for a general number $\tilde{n}$ of Monte Carlo iterations per $\hat{P}_i(\tilde{I}_i)$ as in equation 4.6 of section 4.4 we estimate that with two sequences of IBD realizations, the minimum time factor of 0.34 is achieved with $\tilde{n} = 258$ while for one sequence the minimum time factor of 0.75 is achieved with $\tilde{n} = 80$. Because the minimum estimated time factor is significantly lower with two sequences than one, two sequences should be used, and approximately 250 Monte Carlo iterations should be used per $\hat{P}_i(\tilde{I}_i)$.

We also ran a similar analysis for the half-cousins once removed relationship with individuals 97-1 and 109-4. We analyzed chromosome 1 for these individuals, using the same combinations of other parameters as for the preliminary aunt analysis above. Table 5.5 shows the results from 100 runs of the procedure (CPU times are an average over 5 runs). Acceptance probabilities were approximately 32% and the estimated ratio is approximately 0.99.

From the results in table 5.5 we see that of the parameter values considered, 100 iterations per estimate of $\hat{P}_i(\tilde{I}_i)$ with one sequence of IBD realizations gives the best time factor. The estimated time factor as a function of the number $\tilde{n}$ of iterations per estimate $\hat{P}_i(\tilde{I}_i)$ has its minimum at 0.047 with $\tilde{n} = 27$ when two sequences of IBD realizations are used, and 0.038 with $\tilde{n} = 260$ when one sequence is used. Thus it is best to use one common sequence in this case (see section 4.2.5).

Table 5.5: Investigation of choice of parameters for crossover model analysis for half-cousins once removed.

	common sequence		**two sequences**	
	10 MC	**100 MC**	**10 MC**	**100 MC**
observed s.e.	.065	0.021	0.040	0.026
CPU time	31	95	37	101
time factor	0.129	0.042	0.059	0.068

5.4.2 Results

For the pair 40-1 110-2, the estimate of LOD(chi-square model vs. Haldane's model | aunt) is 0.16 with estimated standard error 0.018, and the estimate of LOD(chi-square model vs. Haldane's model | half-sibs) is 0.29 with estimated standard error 0.007. Thus, compared to the LOD scores with Haldane's model, the chi-square model increases both LOD(aunt vs. unrelated) and LOD(half-sibs vs. unrelated), and decreases LOD(aunt vs. half-sibs), however the aunt relationship remains more likely than the half-sib relationship, as shown in table 5.6.

Table 5.6: Results for pair 40-1 110-2.

	estimate	std. error
LOD(aunt vs. unrelated | chi-square model)	34.53	0.02
LOD(aunt vs. unrelated | Haldane's model)	34.36	—
LOD(half-sibs vs. unrelated | chi-square model)	34.03	0.007
LOD(half-sibs vs. unrelated | Haldane's model)	33.75	—
LOD(aunt vs. half-sibs | chi-square model)	0.50	0.02
LOD(aunt vs. half-sibs | Haldane's model)	0.61	—

For the pair 40-1 51-2, the estimate of LOD(chi-square model vs. Haldane's model |

siblings) is -0.087 with estimated standard error 0.066, and the estimate of LOD(chi-square model vs. Haldane's model | grandparent) is -0.0092 with estimated standard error 0.0045. Thus, we can be about 95% confident that LOD(grandparent vs. unrelated) is lower with the chi-square model than with Haldane's model, but we cannot say whether LOD(siblings vs. unrelated) is higher under the chi-square model or under Haldane's model. We cannot be sure that LOD(siblings vs. grandparent) is less under the chi-square model than under Haldane's model, but we can tell that siblings is a more likely relationship for this pair than grandparent-grandchild under either model. Table 5.7 summarizes the results.

Table 5.7: Results for pair 40-1 51-2.

	estimate	std. error
LOD(siblings vs. unrelated | chi-square model)	22.33	0.07
LOD(siblings vs. unrelated | Haldane's model)	22.41	—
LOD(grandparent vs. unrelated | chi-square model)	19.11	0.007
LOD(grandparent vs. unrelated | Haldane's model)	19.12	—
LOD(siblings vs. grandparent | chi-square model)	3.22	0.07
LOD(siblings vs. grandparent | Haldane's model)	3.29	—

For the pair 97-1 109-4 the estimate of LOD(chi-square model vs. Haldane's model | half-cousins once removed) is -0.017 with estimated standard error 0.008, and the estimate of LOD(chi-square model vs. Haldane's model | second cousins) is -0.010 with estimated standard error 0.021. Thus we can be about 95% confident that LOD(half-cousins once removed vs. unrelated) is lower with the chi-square model than with Haldane's model, but we cannot say whether LOD(second cousins vs. unrelated) is higher under the chi-square model or under Haldane's model. We cannot be sure whether the half-cousins once removed relationship is more likely than the

second cousins relationship under the chi-square model. Table 5.8 summarizes the results.

Table 5.8: Results for pair 97-1 109-4.

	estimate	std. error
LOD(hcr vs. unrelated | chi-square model)	3.04	0.008
LOD(hcr vs. unrelated | Haldane's model)	3.06	—
LOD(sc vs. unrelated | chi-square model)	3.03	0.021
LOD(sc vs. unrelated | Haldane's model)	3.04	—
LOD(hcr vs. sc | chi-square model)	0.012	0.023
LOD(hcr vs. sc | Haldane's model)	0.019	—

5.4.3 Discussion of results

This analysis with the chi-square ($M = 2$) model was most successful with pair 40-1 110-2, for which we were able to tell that the chi-square partly closed the gap between the likelihoods of the aunt and half-sibs relationships. For the pair 40-1 51-2, we could tell that siblings remained a more likely relationship than grandparent-grandchild under the chi-square model, however, due to the high standard error for the estimate of LOD(chi-square model vs. Haldane's model | siblings), we were unable to tell whether LOD(siblings vs. grandparent) is less under the chi-square model than under Haldane's model. For the pair 97-1 109-4 we could not tell whether LOD(second cousins vs. unrelated) is higher under the chi-square model or under Haldane's model because of the high standard error for the estimate of LOD(chi-square model vs. Haldane's model | second cousins). Thus, to obtain more useful results, we need more than the one million MCMC estimates used in this analysis to estimate some of the LOD scores. Alternatively there may be importance sampling improvements that could decrease the standard errors.

Chapter 6

CONCLUSION

6.1 Discussion

Identity by descent and identity by state data contain information about the relationship between a pair of individuals, and about the nature of the underlying process of crossing-over. By calculating likelihoods we can extract all the information. In this thesis, we have developed Monte Carlo methods for calculating such likelihoods, both for continuous IBD data and for discrete IBS data.

6.1.1 Main achievements of this thesis

These methods make possible a range of analyses that were not previously feasible.

Firstly, we can now calculate likelihoods of relationship from continuous IBD data. Continuous IBD data are not currently available, but by analyzing simulated continuous IBD data we can calculate power to distinguish between relationships in the limit as discrete IBS data become more dense and informative. This will allow researchers who are not able to distinguish between relationships using IBS data at current marker density to decide whether it is likely to be worthwhile to have their data genotyped at further genetic markers. If power to distinguish between the relationships is high with continuous IBD data, it may well be worthwhile to undertake further genotyping, whereas if power is low even with continuous IBD data, it will not be possible to distinguish between the relationships even if very densely spaced markers are used.

Secondly, we can work with a range of models for the crossing-over process, in-

cluding models that realistically model interference in the crossing-over process, and models that satisfy the biological requirement of at least one chiasma per chromatid bundle. Incorporating these models into relationship inference will result in improved accuracy and greater power to distinguish between relationships.

Thirdly, we can use IBS and IBD data to distinguish between models for crossing-over, using data that are more readily available than the tetrad and meiosis data that have previously been used to compare models. We are also unconstrained in the number of markers that can be analyzed, which increases the amount of information available to distinguish between models, and allows analysis of models with complicated interference patterns that may be indistinguishable from simpler models when only a handful of markers are used.

Fourthly, we can work with discrete IBS data, with essentially no limit on pedigree size, number of markers or choice of model for the crossing-over process. In combination with existing deterministic methods, this MCMC approach gives great flexibility in analysis of IBS data.

6.1.2 Summary of results

In chapter 2, we presented a Monte Carlo approach to calculation of likelihoods from continuous IBD data. The method is very general in that any reasonable model for the crossing-over process can be used. To illustrate one use of this method, we looked at power to distinguish between relationships with the same expected IBD proportion. We saw that features of the IBD process other than the proportion of IBD do contain information about the relationship, and that in some cases it is possible to distinguish between relationships with the same expected proportion of IBD using the amount of data contained in the human genome.

We also presented two other approaches to calculating likelihoods. The first is an extension to Boehnke and Cox's [3] method and is only valid for discrete data under Haldane's model. The second is a numerical method for half-sib-type relationships.

In chapter 3, we discussed a number of the models for crossing-over that can be found in the literature. For several of these models we were able to describe simplifications to the Monte Carlo procedure which result in reduced computing time. In particular, the method simplifies to that in Browning [6] for Haldane's model, and locations of crossovers do not need to be sampled with this simplification. We also do not need to sample locations of crossovers (except perhaps for the end region, depending on choice of P^*) for the chi-square and truncated Poisson models.

To illustrate use of the Monte Carlo method in crossing-over model inference, we presented an example of distinguishing between models using data from half-sibs. We saw that the Kosambi model is very similar to the chi-square model with parameter $M = 2$, so that the models are quite difficult to distinguish. The truncated Poisson model becomes more similar to the Poisson model as chromosome length increases, so that these models are also difficult to distinguish between for moderately long chromosomes.

At the end of the chapter we discussed the incorporation of sex-specific rates of recombination into the Monte Carlo procedure. In principle this is not at all difficult. However, currently available sex-specific maps are not based on many meioses and hence are not very accurate, so they may not contribute much to current analyses. In section 6.2.1 we outline ideas for incorporating map uncertainty into the analysis. With map uncertainty accounted for, the lack of accuracy in sex-specific maps would be less of a problem.

In chapter 4 we described how the Monte Carlo method for continuous IBD data can be extended to a Markov chain Monte Carlo method for discrete IBS data. The method uses Markov chain Monte Carlo to integrate over possible continuous IBD processes that could underlie the observed discrete IBS data. Unlike the extension to Boehnke and Cox's method, we are not restricted to Haldane's model and can incorporate any kind of model for crossing-over into the analysis. Use of the method was illustrated by analysis of a set of real data in chapter 5.

The data consist of marker genotypes on one hundred Mennonite individuals arranged into small families. Because the individuals come from a small closed community, we expect to see further relationships between pairs of individuals who are reported to be unrelated. Relationships were discovered between twenty pairs of individuals using the extension to Boehnke and Cox's method and assuming Haldane's model. Several pairs of individuals were reanalyzed assuming a chi-square model. For these individuals, the change of models had a small effect on the LOD scores of relationship.

6.2 Future work

Two major areas of future work will dramatically extend the usefulness of these methods. The first of these is incorporation of map uncertainty into the analysis. The second is extension of the method to allow for joint analysis on more than two individuals at a time, and to allow for analysis of some more complicated relationships that we refer to as 'overlapping' bilateral relationships in section 4.3.

6.2.1 Map uncertainty

In this thesis, we have assumed that the genetic distance between any two loci on a chromosome is known. However these distances are merely estimated using the number of observed recombination events between the loci in a collection of meioses. For example, the maps described in Broman et. al. [5], that were used in the analysis in chapter 5, are based on 188 meioses. Error in the maps will reduce power both to distinguish between relationships and to distinguish between models for crossing-over. Without accurate sex-specific maps it may not be worthwhile to include sex-specific rates of crossing-over into analyses.

Given the data from which the maps were estimated, it should be possible to calculate the probability of the estimated map given any possible true underlying

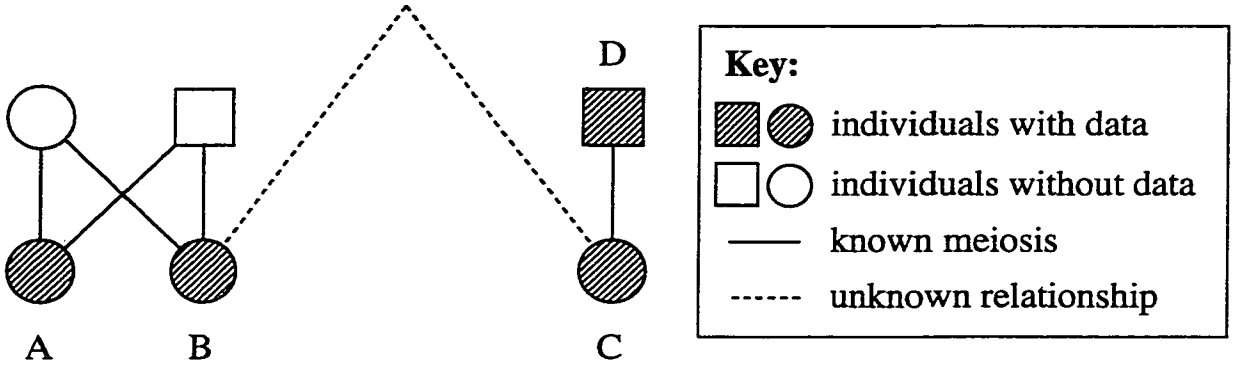

Figure 6.1: Data on multiple individuals.

map. These probabilities could then be incorporated into the analysis, for example by using Monte Carlo to integrate over possible true genetic maps.

With the incorporation of map uncertainty into the likelihood calculation methods, it will be possible to include sex-specific rates of crossing-over into the analysis, and to provide more accurate estimates of relationship and crossing-over model, with better understanding of the variability of those estimates.

6.2.2 Multiple individuals

Although we are often interested in determining the relationship between two individuals, we often also have data on relatives of the two individuals. For example, in figure 6.1 we want to determine the relationship between individuals B and C, but we also have data on individuals A and D who are known to be the sister of B and father of C respectively. When we ignore this secondary data we lose information that could be used in inferring the relationships.

As an example of the additional information in data from relatives, knowing parents' genotypes will often enable haplotyping of the parents' children. Individuals have two copies of each gene, and the two copies lie on different chromosome copies. Haplotyping is the assignment of each of the two genes for each marker to a chro-

mosome. If the data are haplotyped, and the relationships under consideration is unilateral, we need only consider one haplotype (chromosome copy with one gene copy per marker) at a time, which considerably reduces noise and aids in inference. Although data on relatives will not usually result in full haplotyping, the partial haplotyping information that the data give can be included in the analysis.

To perform a full analysis incorporating data from multiple individuals, it may be possible to employ a strategy somewhat like that used for non-overlapping bilateral relationships in section 4.3. That is, it may be possible to partition the combined crossover process into several IBD processes on which moves can be proposed separately, with the processes being brought together in calculating the probability of the observed data given the IBD process realizations.

Once we have a method for analyzing data on multiple individuals, it should be possible to also use the method to analyze overlapping bilateral relationships.

A method for calculating likelihoods on multiple individuals has implications beyond relationship and crossover model inference. Calculation of likelihoods could also be used for gene mapping, in which the crossing-over model and relationships are assumed to be known and the objective is to find the most likely locations on chromosomes of genes responsible for diseases and other characteristics.

BIBLIOGRAPHY

[1] D. Aldous. *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, New York, 1989.

[2] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In O. Shisha, editor, *Inequalities-III: Proceedings of the Third Symposium on Inequalities Held at The University of California, Los Angeles, September 1-9, 1969*, pages 1–8. Academic Press, 1972.

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

[4] E. Borel. *Les Probabilités et la Vie.* Presses Universitaires de France, Paris, sixth edition, 1967.

[5] K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *American Journal of Human Genetics*, 63:861–869, 1998.

[6] S. Browning. Relationship information contained in gamete identity by descent data. *Journal of Computational Biology*, 5:323–334, 1998.

[7] K. P. Donnelly. The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, 23:34–63, 1983.

[8] E. Foss, R. Lande, F. W. Stahl, and C. M. Steinberg. Chiasma interference as a function of genetic distance. *Genetics*, 133:681–691, 1993.

[9] E. J. Foss and F. W. Stahl. A test of a counting model for chiasma interference. *Genetics*, 139:1201–1209, 1995.

[10] C. J. Geyer, O. A. Ryder, L. G. Chemnick, and E. A. Thompson. Analysis of relatedness in the California condors, from DNA fingerprints. *Molecular Biology and Evolution*, 10:571–589, 1993.

[11] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[12] J. B. S. Haldane. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics*, 8:299–309, 1919.

[13] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[14] S. Karlin and U. Liberman. Classifications and comparisons of multilocus recombination distributions. *Proc. Natl. Acad. Sci. USA*, 75:6332–6336, 1978.

[15] T. R. King, W. F. Dove, J. Guénet, B. G. Herrmann, and A. Shedlovsky. Meiotic mapping of murine chromosome 17: The string of loci around *l(17)-2Pas*. *Mammalian Genome*, 1:37–46, 1991.

[16] D. D. Kosambi. The estimation of map distances from recombination values. *Annals of Eugenics*, 12:172–175, 1944.

[17] K. Lange. *Mathematical and statistical methods for genetic analysis*, chapter 12, Models of recombination, pages 206–227. Springer, New York, 1997.

[18] Hammersley J. M. and D. C. Handscomb. *Monte Carlo Methods.* Methuen, London, 1964.

[19] M. S. McPeek and T. P. Speed. Modeling interference in genetic recombination. *Genetics*, 139:1031–1044, 1995.

[20] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, New York, third edition, 1974.

[21] J. Ott. *Analysis of Human Genetic Linkage*, chapter 6, Multipoint linkage analysis. John Hopkins University Press, Baltimore, 1985.

[22] T. P. Speed. What is a genetic map function? In T. P. Speed and M. S. Waterman, editors, *Genetic Mapping and DNA Sequencing*, volume 81 of *IMA Volumes in Mathematics and its Applications*, pages 65–88. Springer-Verlag, New York, 1996.

[23] E. Sturt. A mapping function for human chromosomes. *Annals of Human Genetics*, 40:147–163, 1976.

[24] E. A. Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39:173–188, 1975.

[25] E. A. Thompson. Two-locus and three-locus gene identity by descent in pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology*, 5:261–279, 1988.

[26] E. M. Wijsman and C. I. Amos. Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: Summary of GAW10 contributions. *Genetic Epidemiology*, 14:719–735, 1997.

[27] H. Zhao, M. S. McPeek, and T. P. Speed. Statistical analysis of chromatid interference. *Genetics*, 139:1057–1065, 1995.

[28] H. Zhao and T. P. Speed. On genetic map functions. *Genetics*, 142:1369–1377, 1996.

[29] H. Zhao, T. P. Speed, and M. S. McPeek. Statistical analysis of crossover interference using the chi-square model. *Genetics*, 139:1045–1056, 1995.

Appendix A

DETAILS OF NUMERICAL METHOD FOR HALF-SIB TYPE PEDIGREES

For a half-sib type relationship with k meioses (examples are illustrated in figure A.1), the probability that the two individuals are IBD at each of $n+1$ loci separated by recombination fractions of $u_1, \ldots, u_n$ is $2^{-(k-1)} \prod_{i=1}^{n} (1-u_i)^{k-2}(u_i^2+(1-u_i)^2)$ assuming Haldane's mapping function. (Note that the probability that the two individuals are IBD at each of $n+1$ loci is the $n+1$-locus coefficient of kinship for the individual's related parents.) If x is a genetic distance with recombination fraction r, then $r = \frac{1}{2}(1 - e^{-2x})$ [12]. Hence if $n+1$ loci are equally spaced along a chromosome of length l, the probability that the individuals are IBD for all $n+1$ loci is $2^{-(k-1)}(\frac{1}{2} + \frac{1}{2}e^{-2l/n})^{(k-2)n}(\frac{1}{2} + \frac{1}{2}e^{-4l/n})^{n}$.

Probabilities for combinations of IBD/non-IBD states can be found by appropriate inclusion/exclusion probabilities. In particular define

$$f_\delta(i) = P[\tilde{I}(\delta, 2\delta, \ldots, (i-1)\delta), I(i\delta)|I(0)]$$

where $I(x)$ denotes IBD at point x (distance x Morgans along the chromosome) and $\tilde{I}(x)$ denotes non-IBD at x. Then $f_\delta(i)$ is given recursively by

$$f_\delta(i) = P[I(i\delta)|I(0)] - \sum_{j=1}^{i-1} f_\delta(j) P[I(i\delta)|I(j\delta)].$$

For a half-sib type relationship $P[I(y)|I(x)] = (\frac{1}{2} + \frac{1}{2}e^{-2(y-x)})^{k-2}(\frac{1}{2} + \frac{1}{2}e^{-4(y-x)})$.

Since the IBD process is a renewal process for half-sib type pedigrees so that the lengths of the IBD/non-IBD regions are independent, one can apply $f_\delta(i)$ separately to each non-IBD region (but treating the end regions slightly differently) and then

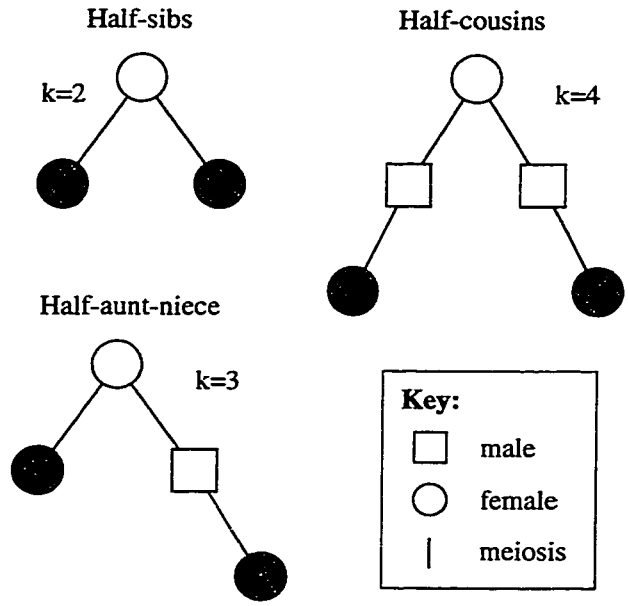

Figure A.1: Examples of half-sib type pedigrees.

multiply in exponential terms for the IBD regions. The $f_\delta(i)$, $i = 1, \ldots, n$ need only be calculated once for each pedigree and δ and sufficiently large n and then stored for further use.

Appendix B

PROBABILITIES FOR STURT AND TRUNCATED POISSON MODELS

In section B.1 we show that the truncated Poisson process is a renewal process while the Sturt process is not. Moreover we show that for a given value of p_0, there is only one count-location chiasma process that is also a renewal process. In section B.2 we obtain probabilities for the crossover process corresponding to the chiasma process for the truncated Poisson and Sturt models.

B.1 Chiasma process probabilities for count-location models

Write $p_n = P(N = n)$ and let X_1 be the location of the first (leftmost) chiasma (if there is no chiasma on the chromosome, we consider X_1 to be greater than L) and let X_2 be the distance between the first chiasma and the second ($X_2 > L - X_1$ if there is only one chiasma).

For $x < L$, noting that $P(X_1 \leq x|N = n) = 1 - \left(1 - \frac{x}{L}\right)^n$ for count-location models,

$$\begin{aligned} P(X_1 \leq x) &= \sum_{n=0}^{\infty} p_n P(X_1 \leq x|N = n) \\ &= \sum_{n=1}^{\infty} p_n \left[1 - \left(1 - \frac{x}{L}\right)^n\right] \\ &= 1 - \sum_{n=1}^{\infty} p_n \left(1 - \frac{x}{L}\right)^n \end{aligned}$$

and the density for X_1 is

$$f_{X_1}(x) = \sum_{n=1}^{\infty} \frac{np_n}{L} \left(1 - \frac{x}{L}\right)^{n-1}. \tag{B.1}$$

Now

$$\begin{aligned} P(N=n|X_1=x) &= P(X_1=x|N=n)P(N=n)/P(X_1=x) \\ &= \frac{np_n}{L}\left(1-\frac{x}{L}\right)^{n-1} \Big/ \sum_{m=1}^{\infty}\frac{mp_m}{L}\left(1-\frac{x}{L}\right)^{m-1} \\ &= np_n\left(1-\frac{x}{L}\right)^{n-1} \Big/ \sum_{m=1}^{\infty} mp_m\left(1-\frac{x}{L}\right)^{m-1} \end{aligned}$$

and for $y < L - x$,

$$\begin{aligned} & P(X_2 \le y|X_1=x) \\ &= \sum_{n=0}^{\infty} P(N=n|X_1=x)\, P(X_2 \le y|X_1=x, N=n) \\ &= \sum_{n=2}^{\infty} \frac{np_n\left(1-\frac{x}{L}\right)^{n-1}}{\sum_{m=1}^{\infty} mp_m\left(1-\frac{x}{L}\right)^{m-1}}\left[1-\left(1-\frac{y}{L-x}\right)^{n-1}\right] \\ &= \frac{\sum_{n=2}^{\infty} np_n\left(1-\frac{x}{L}\right)^{n-1}}{\sum_{n=1}^{\infty} np_n\left(1-\frac{x}{L}\right)^{n-1}} - \frac{\sum_{n=2}^{\infty} np_n\left(1-\frac{x}{L}\right)^{n-1}\left(1-\frac{y}{L-x}\right)^{n-1}}{\sum_{n=1}^{\infty} np_n\left(1-\frac{x}{L}\right)^{n-1}} \\ &= 1 - \frac{\sum_{n=1}^{\infty} np_n\left(1-\frac{x}{L}\right)^{n-1}\left(1-\frac{y}{L-x}\right)^{n-1}}{\sum_{n=1}^{\infty} np_n\left(1-\frac{x}{L}\right)^{n-1}} \\ &= 1 - \frac{\sum_{n=1}^{\infty} np_n\left(1-\frac{x+y}{L}\right)^{n-1}}{\sum_{n=1}^{\infty} np_n\left(1-\frac{x}{L}\right)^{n-1}} \end{aligned} \tag{B.2}$$

B.1.1 The truncated Poisson process is a renewal process

For the truncated Poisson model, $p_0 = 0$ and $p_n = \frac{e^{-\alpha L}(\alpha L)^n}{n!\,(1-e^{-\alpha L})}$ for $n \ge 1$. Hence from equation B.1

$$\begin{aligned} f_{X_1}(x) &= \sum_{n=1}^{\infty} \frac{n\, e^{-\alpha L}(\alpha L)^n}{Ln!\,(1-e^{-\alpha L})}\left(1-\frac{x}{L}\right)^{n-1} \\ &= \frac{\alpha\, e^{-\alpha L}}{1-e^{-\alpha L}} \sum_{m=0}^{\infty} \alpha^m (L-x)^m/m! \\ &= \alpha\, e^{-\alpha L} e^{\alpha(L-x)}/(1-e^{-\alpha L}) \\ &= 2\, e^{-\alpha x} \qquad\qquad \text{since } \alpha/2 = 1-e^{-\alpha L} \end{aligned}$$

and from equation B.2

$$\begin{aligned} P(X_2 \leq y | X_1 = x) &= 1 - \frac{\sum_{n=1}^{\infty} (\alpha L)^n \left(1 - \frac{x+y}{L}\right)^{n-1} / (n-1)!}{\sum_{n=1}^{\infty} (\alpha L)^n \left(1 - \frac{x}{L}\right)^{n-1} / (n-1)!} \\ &= 1 - \exp\left(\alpha L \left(1 - \frac{x+y}{L}\right) - \alpha L \left(1 - \frac{x}{L}\right)\right) \\ &= 1 - e^{-\alpha y}. \end{aligned}$$

Since this does not depend on x, $P(X_2 \leq y | X_1 = x) = P(X_2 \leq y)$ and the process must be a renewal process. Note that $F_{X_2}(x) = P(X_2 \leq y) = 1 - e^{-\alpha x}$, so that $f_{X_1}(x) = 2\left(1 - F_{X_2}(x)\right)$, as expected for a stationary renewal process of rate 2.

B.1.2 The Sturt process is not a renewal process

For the Sturt model, $p_0 = 0$ and $p_n = e^{-\lambda}\lambda^{n-1}/(n-1)!$ for $n \geq 1$, with $\lambda = 2L - 1$. Noting that

$$\begin{aligned} \sum_{n=1}^{\infty} n e^{-\beta} \gamma^{n-1} / (n-1)! &= \sum_{n=1}^{\infty} e^{-\beta} \gamma^{n-1} / (n-1)! + \sum_{n=1}^{\infty} (n-1) e^{-\beta} \gamma^{n-1} / (n-1)! \\ &= e^{-\beta} e^{\gamma} (1 + \gamma), \end{aligned}$$

and substituting into equation B.2 we find that for $y < L - x$

$$\begin{aligned} P(X_2 \leq y | X_1 = x) &= 1 - \frac{\sum_{n=1}^{\infty} \frac{n e^{-\lambda} \lambda^{n-1}}{(n-1)!} \left(1 - \frac{x+y}{L}\right)^{n-1}}{\sum_{n=1}^{\infty} \frac{n e^{-\lambda} \lambda^{n-1}}{(n-1)!} \left(1 - \frac{x}{L}\right)^{n-1}} \\ &= 1 - \frac{e^{-\lambda} e^{\lambda\left(1 - \frac{x+y}{L}\right)} \left[1 + \lambda \left(1 - \frac{x}{L}\right)\left(1 - \frac{y}{L-x}\right)\right]}{e^{-\lambda} e^{\lambda\left(1 - \frac{x}{L}\right)} \left[1 + \lambda \left(1 - \frac{x}{L}\right)\right]} \\ &= 1 - e^{-\lambda y / L} \left(1 + \lambda (L - x - y)/L\right) / \left(1 + \lambda (L - x)/L\right) \end{aligned}$$

which does depend on x, hence the Sturt process cannot be a stationary renewal process.

B.1.3 The truncated Poisson model is unique in being a count location model and renewal model

For a renewal chiasma model, $f_{X_1}(x) = 2\,(1 - F_{X_2}(x))$. Hence for a count-location chiasma model that is also a renewal model, equations B.1 and B.2 imply that

$$\left(\sum_{n=1}^{\infty} np_n \left(1-\frac{x}{L}\right)^{n-1}\right)\left(\sum_{n=1}^{\infty} np_n \left(1-\frac{x}{L}\right)^{n-1}\right) = 2L\sum_{n=1}^{\infty} np_n \left(1-\frac{x+y}{L}\right)^{n-1}. \quad \text{(B.3)}$$

Write $\phi(s) = \sum_{n=1}^{\infty} np_n(1-s)^{n-1}$. Then equation B.3 can be written

$$\phi(\frac{x}{L})\phi(\frac{y}{L}) = 2L\phi(\frac{x+y}{L})$$

and this holds for all $\{(x,y) : x \geq 0, y \geq 0, x+y \leq L\}$. Hence,

$$\phi(s)\phi(t) = 2L\phi(s+t) \quad \text{(B.4)}$$

for all $\{(s,t) : s \geq 0, t \geq 0, s+t \leq 1\}$. Now $\phi(1) = E(N) = 2L$ by the definition of genetic distance, and $\phi(0) = \lim_{s \searrow 0} \phi(s) = p_0$.

Suppose p_0 is set at some value. Then we can show that ϕ (and hence $\{p_n, n = 1, 2, \ldots\}$) is uniquely determined. To show this by induction, supposed that for some n the value of $\phi(\frac{i}{2^n})$ is determined for $i = 0, 1, \ldots, 2^n$ (for fixed L and p_0 this holds for $n = 0$). Then trivially $\phi(\frac{0}{2^{n+1}}) = \phi(\frac{0}{2^n})$ which is determined. From equation B.4, $\phi(\frac{1}{2^{n+1}})\phi(\frac{1}{2^{n+1}}) = 2L\phi(\frac{1}{2^n})$ thus $\phi(\frac{1}{2^{n+1}}) = \sqrt{2L\phi(\frac{1}{2^n})}$ which is also determined. Further, $\phi(\frac{i}{2^{n+1}}) = (\frac{1}{2L})^{i-1}\phi^i(\frac{1}{2^{n+1}})$ is determined for $i = 2, 3, \ldots, 2^{n+1}$. Thus by induction $\phi(\frac{i}{2^n})$ is determined for all n and $i = 0, 1, \ldots, 2^n$. By continuity of ϕ, $\phi(s)$ is determined for all $0 \leq s \leq 1$.

Hence for any fixed value of p_0, there is only one count-location chiasma process that is also a renewal process. For $p_0 = 0$, this unique process is the truncated Poisson process. For $p_0 = e^{-2L}$ the unique renewal count-location chiasma process is the Poisson process (Haldane's model).

B.2 Crossover process probabilities

Assuming NCI (no chromatid interference, see section 3.2), the crossover process corresponding to a count-location chiasma process is also a count-location process. If N is the number of chiasmata on a chiasma process realization and M is the number of crossovers on the corresponding crossover process realization, the distribution of M given N is binomial:

$$P(M = m|N = n) = \frac{n!}{m!\,(n-m)!}\frac{1}{2^n}$$

and hence

$$P(M = m) = \sum_{n=m}^{\infty} \frac{n!}{m!\,(n-m)!}\frac{1}{2^n}P(N = n).$$

For the truncated Poisson,

$$\begin{aligned} P(M = m) &= \sum_{n=m}^{\infty} \frac{e^{-\alpha L}(\alpha L)^n}{m!\,(n-m)!\,(1-e^{-\alpha L})\,2^n} \\ &= \frac{e^{-\alpha L/2}(\alpha L/2)^m}{m!\,(1-e^{-\alpha L})} \qquad \text{for } m \geq 1 \\ \text{and } P(M = 0) &= \sum_{n=1}^{\infty} \frac{e^{-\alpha L}(\alpha L)^n}{m!\,(1-e^{-\alpha L})\,2^n} \\ &= \frac{e^{-\alpha L}}{1-e^{-\alpha L}}(e^{\alpha L/2}-1). \end{aligned}$$

By thinning, the truncated Poisson crossover process is a renewal process and has renewal density $g_{X_2}(x) = \alpha e^{-\alpha x/2}/2$ for $x < L$.

For the Sturt model,

$$\begin{aligned} P(M = m) &= \sum_{n=m}^{\infty} \frac{n\,e^{-\lambda}\lambda^{n-1}}{m!\,(n-m)!\,2^n} \\ &= \frac{1}{2}\left[\frac{e^{-\lambda/2}(\lambda/2)^{m-1}}{(m-1)!} + \frac{e^{-\lambda/2}(\lambda/2)^m}{m!}\right] \qquad \text{for } m \geq 1 \\ \text{and } P(M = 0) &= \sum_{n=1}^{\infty} \frac{e^{-\lambda}\lambda^{n-1}}{(n-1)!\,2^n} \\ &= e^{-\lambda/2}/2. \end{aligned}$$

Appendix C

SEX-SPECIFIC MAP METHOD FOR HALDANE'S MODEL

In this appendix we show how sex-specific (or individual-specific) rates of crossing-over can be incorporated into the Monte Carlo method of Browning [6]. The method is described in the context of Haldane's model, but also applies to the chi-square model as discussed in chapter 3.

C.1 The Poisson process model for sex-specific rates

Suppose we have some base map — such as a sex-averaged map. Define genetic length so that the base map has crossovers occurring at rate one (per Morgan). For each meiosis, we assume that crossovers occur as an inhomogeneous Poisson process relative to the base map. That is, crossovers occur as a memoryless process but some sections of chromosomes may be more prone to crossing over than others (compared to the base map or to other meioses). Let $\lambda_i(t)$ be the crossover rate at location t for meiosis i, with location measured in Morgans on the base map.

If we consider sex difference in rates of crossing over, there are two rate functions, λ_f and λ_m, for meioses from females and males respectively. (The methods given can also be applied to individual rate functions, if sufficient information is available to define these.) We will assume that the base map is the sex-averaged map, since this will most often be the case, but with simple modifications the method given below can be used with other base maps. Given this assumption, a combined crossover process for one male and one female meiosis will have rate 2, that is, $\lambda_f(t) + \lambda_m(t) = 2$ for all

locations t on the genome.

C.2 The pedigree-average map

Calculations and concepts will be significantly simplified if we switch to a pedigree-average map. A pedigree-average map, like the sex-average map, defines distance so that the *average* distance between adjacent crossovers is 1. Whereas the sex-average map averages equal numbers of male and female meioses, a pedigree-average map averages over the meioses in a given pedigree.

Before we can describe the transformation to a pedigree-average map, we need to present the relevant probability distributions. Starting from some position t^*, the distance X_i to the next crossover event on meiosis i has distribution function [reference to a book with this equation]

$$P(X_i \leq x) = 1 - \exp\left\{-\int_{t^*}^{t^*+x} \lambda_i(t)\, dt\right\}. \tag{C.1}$$

Let k be the number of meioses in the pedigree. The k crossover processes are independent of one another, so the probability distribution function of the distance $X = \min\{X_i\}$ to the next crossover event on the combined crossover process is

$$\begin{aligned} P(X \leq x) &= 1 - \prod_{i=1}^{k} P(X_i > x) \\ &= 1 - \exp\left\{-\int_{t^*}^{t^*+x} \sum_{i=1}^{k} \lambda_i(t)\, dt\right\}. \end{aligned} \tag{C.2}$$

(Note that if there are f meioses from females in the pedigree, $\sum_{i=1}^{k} \lambda_i(t)$ is equal to $f\,\lambda_f(t) + (k-f)\,\lambda_m(t)$.)

Now we can define the transformation function α that maps a location on the sex-average map to the corresponding location on the pedigree-average map:

$$\alpha(t) = \int_0^t \sum_{i=1}^{k} \lambda_i(x)/k\, dx.$$

Also, the rate function for meiosis i relative to the pedigree-average map is

$$\mu_i(s) = \frac{\lambda_i(\alpha^{-1}(s))}{\sum_{i=1}^{k} \lambda_i(\alpha^{-1}(s))/k} \tag{C.3}$$

where α^{-1} is the inverse transformation mapping locations on the pedigree-average map back to the corresponding locations on the sex-average map.

Starting from a position s^* on the pedigree-average map corresponding to $t^* = \alpha^{-1}(s^*)$ on the sex-average map, the pedigree-average-map distance Y_i (corresponding to the sex-average-map distance X_i defined above equation (C.1)) to the next crossover event on meiosis i has distribution function

$$\begin{aligned} P(Y_i \le y) &= P\left(X_i \le \alpha^{-1}(y + s^*) - \alpha^{-1}(s^*)\right) \\ &= 1 - \exp\left\{-\int_{\alpha^{-1}(s^*)}^{\alpha^{-1}(s^*+y)} \lambda_i(t)\, dt\right\}. \end{aligned}$$

By change of variables $s = \alpha(t)$, this becomes

$$\begin{aligned} P(Y_i \le y) &= 1 - \exp\left\{-\int_{s^*}^{s^*+y} \frac{\lambda_i(\alpha^{-1}(s))}{\sum_{i=1}^{k} \lambda_i(\alpha^{-1}(s))/k}\, ds\right\} \\ &= 1 - \exp\left\{-\int_{s^*}^{s^*+y} \mu_i(s)\, ds\right\} \end{aligned} \tag{C.4}$$

which is what is expected (c.f. equation (C.1)).

The distribution function of the pedigree-average-map distance Y to the next crossover event on the combined crossover process (the minimum of the Y_i's) is (see equation (C.2))

$$P(Y \le y) = 1 - \exp\left\{-\int_{s^*}^{s^*+y} \sum_{i=1}^{k} \mu_i(s)\, ds\right\}.$$

By summing equation (C.3) over all k meioses, we find that $\sum_{i=1}^{k} \mu_i(s) = k$ (this can also be seen directly — since the pedigree-average map has an average *distance* of one between crossovers, the average *rate* of crossing over must also equal one). Hence

$$P(Y \le y) = 1 - \exp(-ky) \tag{C.5}$$

and Y has an exponential distribution with rate k regardless of the starting location s^*. This attribute is the reason for transforming to the pedigree-average map.

C.3 Monte Carlo simulation

We will describe here a method for simulating the combined crossover process using sex-specific genetic maps. Simulated processes can be converted into simulated data, by recording locations of change in IBD status only, or into simulated jump chains, by recording the series of states but not the locations.

This method simulates realizations of the combined crossover process relative to the pedigree-average map. Once a realization has been simulated, it can be transformed to the sex-average-map by application of α^{-1}.

Suppose a realization has been simulated up to location s^* (on the pedigree-average map). The process is memoryless since each crossover process is assumed to be memoryless and the crossover processes for the k meioses are independent of each other, so it does not matter what the current state of the process is. The distance Y to the next crossover event will be exponentially distributed with rate k (by equation (C.5)), and the probability that the crossover occurs on meiosis j given that $Y = y$ is

$$\begin{aligned} P(Y_j = \min\{Y_i\} | \min\{Y_i\} = y) &= \frac{P(Y_j = y, \text{ and } Y_i > y \text{ for all } i \neq j)}{P(Y = y)} \\ &= \frac{P(Y_j = y) \prod_{i \neq j} P(Y_i > y)}{P(Y = y)} \end{aligned}$$

since the Y_i's are independent. Here P is used to represent the probability measure that suits the context, which is at times a probability distribution function, a probability density function, or a hybrid. Substituting the probability distribution functions and their derivatives from equations (C.4) and (C.5),

$$\begin{aligned} & P(Y_j = \min\{Y_i\} | \min\{Y_i\} = y) \\ &= \frac{\mu_j(s^* + y) \, \exp\left\{ - \int_{s^*}^{s^*+y} \mu_j(s) \, ds \right\} \prod_{i \neq j} \exp\left\{ - \int_{s^*}^{s^*+y} \mu_i(s) \, ds \right\}}{k \, \exp(-ky)} \\ &= \frac{\mu_j(s^* + y) \, \exp\left\{ - \int_{s^*}^{s^*+y} \sum_{i=1}^{k} \mu_i(s) \, ds \right\}}{k \, \exp(-ky)} \\ &= \mu_j(s^* + y) \, / k. \end{aligned} \tag{C.6}$$

Hence the realization can be continued by generating a distance y to the next crossover event of the realization from the exponential distribution with rate k, and then selecting an element of the inheritance vector to change according to the probabilities given in equation (C.6).

C.4 Monte Carlo likelihoods

As in chapter 2 we wish to calculate the likelihood $L(\text{model}|\mathbf{I})$ of a model given a chromosome $\mathbf{I}$ of IBD data, which is just the probability $P(\mathbf{I})$ of the data given the model. This probability is equal to

$$P(\mathbf{I}) = \sum_{\text{chains}} P(\mathbf{I}|\text{chain})P(\text{chain})$$

and can be estimated by the average

$$\hat{P}(\mathbf{I}) = \sum_{j=1}^{n} P(\mathbf{I}|\text{chain } j)/n$$

where the n jump chains are simulated from $P(\text{chain})$, the distribution of jump chains under the model. Section C.3 describes an algorithm for simulating these jump chains.

Now we need to calculate the probability $P(\mathbf{I}|\text{chain})$. Consider a chain that has the same IBD status at the first region as do the data (otherwise the probability is zero). Following figure 1.4 we label the locations of the switches between IBD and non-IBD regions of the data $\mathbf{I} = (0 = l_0 < l_1 < \ldots < l_m = L)$ from left to right. The numbers of steps of the chain in each region are $\mathbf{c} = (c_1, c_2, \ldots, c_m)$, and the sequence of meiosis indicators that were changed at each step in region i are labeled $\mathbf{s_i} = (s_{i1}, s_{i2}, \ldots, s_{ic_i})$ for $i = 1, 2, \ldots, m$.

The combined crossover process is a hidden Markov model for the data, hence, given the sequence $\mathbf{s_i}$ of steps of the chain corresponding to region i of the data and the location l_{i-1} of the previous switch of IBD/non-IBD status, the location l_i of the ith switch is independent of the steps of the chain outside the region and the locations

of the other switches in IBD status of the data. Hence,

$$P(\mathbf{I}|\text{chain}) = \prod_{i=1}^{m} P(l_i|l_{i-1}, \mathbf{s_i}) \tag{C.7}$$

and we can consider one region at a time.

Let us consider any region of the data, except the rightmost region. The chain itself does not have locations associated with it, but we can calculate $P(l_i|l_{i-1}, \mathbf{s_i})$ by integrating over locations for the steps of the chain in this region that are consistent with the data. Let A_i be the set of possible locations: $A_i = \{\mathbf{t_i} : l_{i-1} < t_{i1} < t_{i2} < \ldots < t_{ic_i} = l_i\}$. Then

$$P(l_i|l_{i-1}, \mathbf{s_i}) = \int_{A_i} P(l_i, \mathbf{t_i}|l_{i-1}, \mathbf{s_i})\, d\mathbf{t_i}. \tag{C.8}$$

Let $\lambda(A_i)$ be the Lebesgue measure of A_i, which is equal to

$$\int_{A_i} d\mathbf{t} = \frac{(l_i - l_{i-1})^{c_i - 1}}{(c_i - 1)!}.$$

If a vector $\mathbf{t_i^*}$ is chosen uniformly at random from A_i then

$$\hat{P}(l_i|l_{i-1}, \mathbf{s_i}) = \lambda(A_i) P(l_i, \mathbf{t_i^*}|l_{i-1}, \mathbf{s_i}) \tag{C.9}$$

is an unbiased estimate of $P(l_i|l_{i-1}\mathbf{s_i})$ and can be used in its place in the Monte Carlo likelihood estimate. Alternatively, one can sample several values of $\mathbf{t_i}$ and use an average.

Let $\tilde{l}_m$ be the sampled location of the final step of the chain. Calculation of $P(\tilde{l}_m > L|l_{m-1}, \mathbf{s_m})$, for the rightmost region of the data, is similar except that the location $\tilde{l}_m$ does not correspond to a change in IBD status, so the set of possible locations for steps of the chain is $A_m = \{\mathbf{t_m} : l_{m-1} < t_{m1} < t_{m2} < \ldots < t_{mc_m}$ and $t_{mc_m} = \tilde{l}_m > L\}$. We cannot sample $\mathbf{t_m}$ from a uniform distribution since $t_{mc_m} = \tilde{l}_m$ doesn't have a fixed value. Instead we choose some distribution $P^*(\mathbf{t_m}|t_{mc_m} > L)$ from which to sample. For example we could choose to sample $\mathbf{t_m}$ so that distances between adjacent

crossovers have an exponential distribution with rate k conditional on $t_{mc_m} > L$, in which case

$$\begin{aligned} P^*(\mathbf{t_m}|t_{mc_m} > L) &= \frac{\prod_{j=1}^{c_m} k e^{-k(t_{mj}-t_{m,j-1})}}{\int_{L-l_{m-1}}^{\infty} x^{c_m-1}e^{-kx}k^{c_m}/(c_m-1)!\,dx} \\ &= \frac{k^{c_m}e^{-k(t_{mc_m}-l_{m-1})}}{\sum_{j=0}^{c_m-1}(l_m-l_{m-1})^j e^{-k(L-l_{m-1})}k^j/j!}. \end{aligned}$$

Now we can write

$$P(\tilde{l}_m > L|l_{m-1}, \mathbf{s_m}) = \int_{A_m} \frac{P(\tilde{l}_m > L, \mathbf{t_m}|l_{m-1}, \mathbf{s_m})}{P^*(\mathbf{t_m}|t_{mc_m} > L)} P^*(\mathbf{t_m}|t_{mc_m} > L)\, d\mathbf{t_m} \qquad \text{(C.10)}$$

and sample a vector t_m^* from P^*. Then

$$\hat{P}(\tilde{l}_m > L|l_{m-1}, \mathbf{s_m}) = P(\tilde{l}_m > L, \mathbf{t_m^*}|l_{m-1}, \mathbf{s_m})/P^*(\mathbf{t_m}|t_{mc_m} > l) \qquad \text{(C.11)}$$

is an unbiased estimate of $P(\tilde{l}_m > L|l_{m-1}, \mathbf{s_m})$.

The probability $P(l_i, \mathbf{t_i}|dli-1, \mathbf{s_i})$ in the equations above can also be written as

$$\begin{aligned} &P(l_i, \mathbf{t_i}|l_{i-1}, \mathbf{s_i}) \\ &\quad = P(\text{crossovers at } t_{i1}, t_{i2}, \ldots, t_{ic_i}|\text{crossovers on } s_{i1}, s_{i2}, \ldots, s_{ic_i}). \end{aligned}$$

Using the memoryless property of the combined crossover process once again, this probability factors into

$$\begin{aligned} &P(t_{i1}, t_{i2}, \ldots, t_{ic_i}|s_{i1}, s_{i2}, \ldots, s_{ic_i}) \\ &\quad = \prod_{j=1}^{c_i} P(\text{crossover at } t_{ij}|\text{crossover on } s_{ij}, \text{previous crossover at } t_{i,j-1}). \end{aligned}$$

Applying Bayes rule, and noting that $P(s_{ij}|t_{ij}, t_{i,j-1}) = P(s_{ij}|t_{ij})$,

$$\begin{aligned} P(t_{ij}|s_{ij}, t_{i,j-1}) &= \frac{P(s_{ij}|t_{ij}, t_{i,j-1})P(t_{ij}|t_{i,j-1})}{\int_{t_{i,j-1}}^{\infty} P(s_{ij}|t, t_{i,j-1})P(t|t_{i,j-1})dt} \\ &= \frac{P(s_{ij}|t_{ij})P(t_{ij}|t_{i,j-1})}{\int_{t_{i,j-1}}^{\infty} P(s_{ij}|t)P(t|t_{i,j-1})dt} \end{aligned} \qquad \text{(C.12)}$$

with integration in the denominator of equation (C.12) being over possible values of t_{ij}, the location of the jth crossover in the ith region.

From equation (C.5) we know that,

$$P(t_{ij}|t_{i,j-1}) = k \exp(-k(t_{ij} - t_{i,j-1}))$$

and from equation (C.6),

$$P(s_{ij}|t_{ij}) = \mu_{s_{ij}}(t_{ij})/k \tag{C.13}$$

so that

$$P(l_i, \mathbf{t_i}|l_{i-1}, \mathbf{s_i}) = \frac{e^{-kL} \prod_{j=1}^{c_i} \mu_{s_{ij}}(t_{ij})}{\prod_{j=1}^{c_i} \int_{t_{i,j-1}}^{\infty} \mu_{s_{ij}}(t) e^{-k(t-t_{i,j-1})} dt} \tag{C.14}$$

Putting the pieces in equations (C.7) to (C.14) together, an unbiased estimate of $P(\mathbf{I}|\text{chain})$ is

$$\begin{aligned}
&\hat{P}(\mathbf{I}|\text{chain}) \\
&= \prod_{i=1}^{m} \hat{P}(l_i|l_{i-1}, \mathbf{s_i}) \\
&= \left(\prod_{i=1}^{m-1} \lambda(A_i) \prod_{j=1}^{c_i} P(t_{ij}|s_{ij}, t_{i,j-1}) \right) \frac{\prod_{j=1}^{c_m} P(t_{mj}|s_{mj}, t_{m,j-1})}{P^*(\mathbf{t_m}|t_{mc_m} > l)} \\
&= \left(\prod_{i=1}^{m} \frac{e^{-kL} \prod_{j=1}^{c_i} \mu_{s_{ij}}(t_{ij})}{\prod_{j=1}^{c_i} \int_{t_{i,j-1}}^{\infty} \mu_{s_{ij}}(t) e^{-k(t-t_{i,j-1})} dt} \right) \frac{\prod_{i=1}^{m-1} \lambda(A_i)}{P^*(\mathbf{t_m}|t_{mc_m} > l)}
\end{aligned}$$

where the t_{ij} are sampled valued as described above.

Appendix D

APPROXIMATIONS TO THE LIKELIHOOD BASED ON LOCAL PROPERTIES

In this appendix we investigate local properties of the identity by descent process and obtain several methods for approximating relationship likelihoods. These methods can give good approximations to the likelihood with much less computation than the full Monte Carlo likelihood procedure. In addition the methods provide some insight into properties of the identity by descent process.

Simulation studies presented in Browning [6] showed that relationship information along the genome is, in a sense, quite local. Cutting chromosomes into smaller pieces did not result in detectable loss of information.

We will present several methods for approximating the likelihood that are not as computationally intensive as the Monte Carlo likelihood method in Browning. We start by investigating a very simple model, the Markov model, that imposes a lot of structure on the approximation. This approximation is a special case of regression, so we compare the approximation to the outcome of a regression. We then look at some renewal approximations, which assume that the lengths of IBD and non-IBD regions are essentially independent. After trying to fit exponential distributions to the region lengths, we look at the results of fitting empirical region length distributions. The success or otherwise of these methods provides some insight into features of the IBD process.

The usefulness of the approximations is evaluated using 200 simulated chromosomes of data, each of length two Morgans, from the half-first-cousins-once-removed

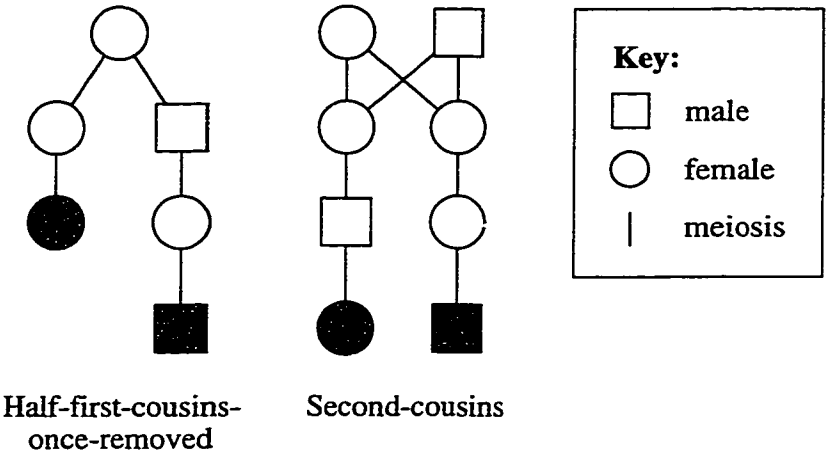

Figure D.1: Pedigrees used in evaluating the approximations.

relationship. The two Morgan length was chosen so that a large number of chromosomes with more than one IBD region would be observed, without having many overly complicated chromosomes with very many IBD regions. The likelihoods of second-cousins and of half-first-cousins-once-removed were calculated for each chromosome using the Monte Carlo approach in Browning [6] with up to three million Monte Carlo iterations. The two relationships are shown in figure D.1. The median Monte Carlo coefficient of variation (standard error of the estimate divided by the estimate) was 0.002 and the highest value was 0.011. These errors are sufficiently small that they will not affect the graphical presentation of results.

D.1 A Markov model

If the IBD process were Markov, relationship information would be completely local. No information would be lost by cutting chromosomes into smaller pieces.

Except for trivial cases, the IBD process is not Markov, but in order to obtain an approximation to the likelihood, we will approximate the process with a Markov model. We model both IBD and non-IBD regions as having independent exponentially

distributed lengths. Under this model the observed IBD proportion and numbers of IBD and non-IBD regions are sufficient statistics for the data.

A natural choice for the rate of the exponential distribution for IBD region lengths is

$$\begin{aligned}\mu &= \lim_{x\to 0} P(\text{non-IBD at x}|\text{IBD at 0})/x \\ &= \lim_{x\to 0}(k_1 - K_2(x))/(k_1 x) \\ &= k_2/k_1\end{aligned}$$

where

$$k_1 = P(\text{IBD at a given locus})$$

is the coefficient of kinship and

$$K_2(x) = P(\text{IBD at 0 and x}) = k_1 - k_2 x + o(x)$$

for some constant k_2 is the two-locus coefficient of kinship. Similarly we will choose

$$\begin{aligned}\lambda &= \lim_{x\to 0} P(\text{IBD at x}|\text{non-IBD at 0})/x \\ &= \lim_{x\to 0}(k_1 - K_2(x))/(1 - k_1)x \\ &= k_2/(1 - k_1)\end{aligned}$$

for the rate of the non-IBD distribution. The mean time that this process is IBD is $\frac{1}{\mu}/(\frac{1}{\mu} + \frac{1}{\lambda}) = k_1$, which is the same as for the true IBD process.

Under this model, the fitted log likelihood is a linear function of IBD proportion, with different intercepts for different numbers of IBD and non-IBD regions. The fit of the model can be compared with a linear regression of the same form.

For the comparison, we use the simulated data described in the introduction, however to avoid having any regression coefficients determined by only one value (and thus inflating the percentage of variation explained by the model) several points have to be discarded. Only one chromosome has 6 IBD regions, so it is discarded.

There is only one chromosome that is IBD at both ends, so it is discarded. The null chromosome which is completely non-IBD has to be discarded; there are 99 of these in the data set. If there had been any chromosomes completely IBD these would also have been discarded.

The regression does a good job of estimating the log likelihoods for each of the two relationships, explaining 94% of the variation in the data in the case of half-first-cousins-once-removed likelihood, and 96% in the case of second-cousins. As can be seen from the figure, using the same slope (coefficient associated with IBD proportion) regardless of numbers of IBD and non-IBD regions seems to fit well. Variation about the model is lowest for chromosomes with only one IBD region, and a good approximation to the likelihood is obtained for these chromosomes. Likelihoods of chromosomes with more than one IBD region are more variable around the fitted line, indicating that the linear model is not adequate for approximating the likelihoods of these chromosomes.

Compared with the regression, the Markov model does not estimate the intercepts (effects of numbers of IBD and non-IBD regions) very well. The direction of error is the same for both the second-cousins and half-first-cousins-once-removed relationships, so that much of the error cancels out when we take the log likelihood ratio.

D.1.1 Markov approximation to the log likelihood when the coefficients of kinship are equal

The expression for the likelihood ratio which results from the Markov model above can be also obtained in the following manner when the coefficient of kinship is the same for both relationships in question. Suppose the information in the data is contained only in the number of switches between IBD and non-IBD regions and in the IBD proportion. Then this information will not be lost if we cut the data into smaller and smaller pieces. The likelihood of each relationship is a function of the size of the pieces, but in the limit (as the pieces become infinitesimally small) the likelihood

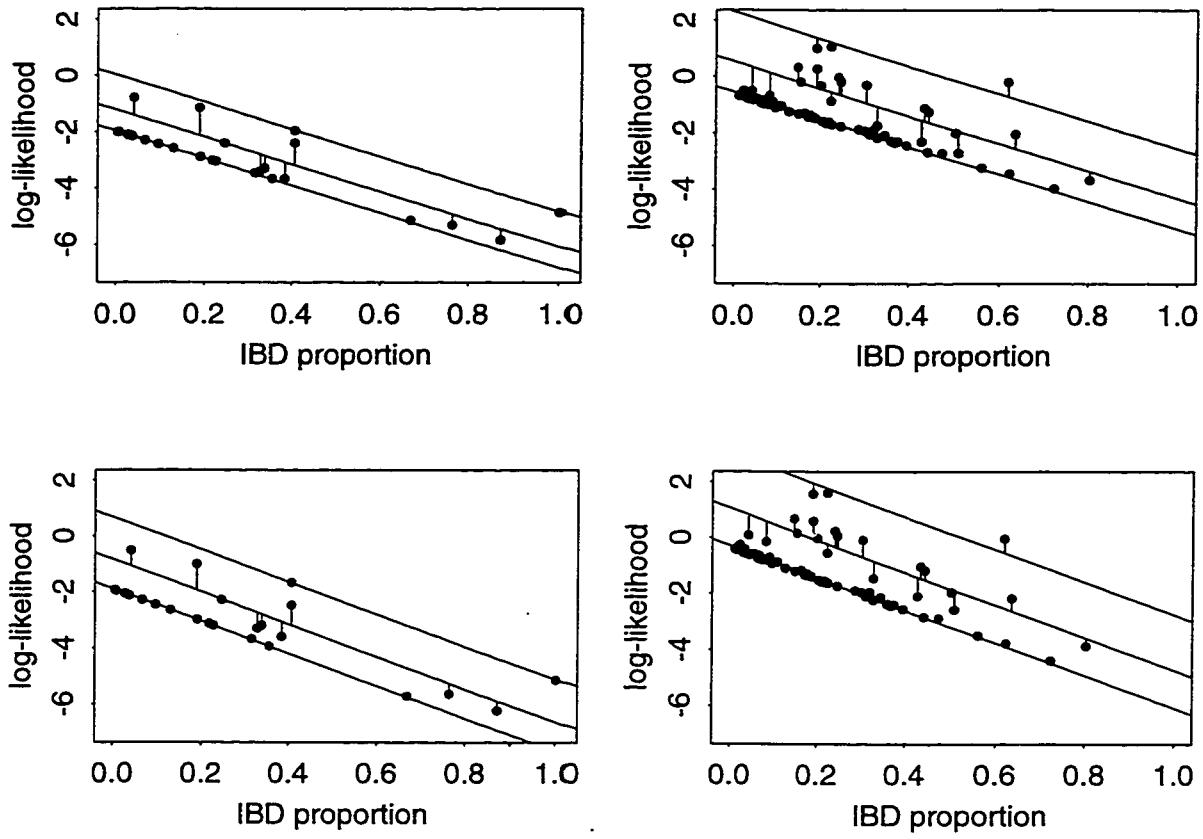

Figure D.2: **Regression approximation.** The fitted lines and deviation of observed log likelihood values from those lines for the regression. The upper row of plots show log likelihood ratios for the half-first-cousins-once-removed relationship, and the lower row of plots show log likelihood ratios for the second cousins relationship. The chromosomes are separated into two plots for each relationship, the left for chromosomes that are IBD at one end and the right for chromosomes non-IBD at both ends, to aid readability. In each plot the three lines are for chromosomes with, from bottom to top, 1, 2 and 3 IBD regions.

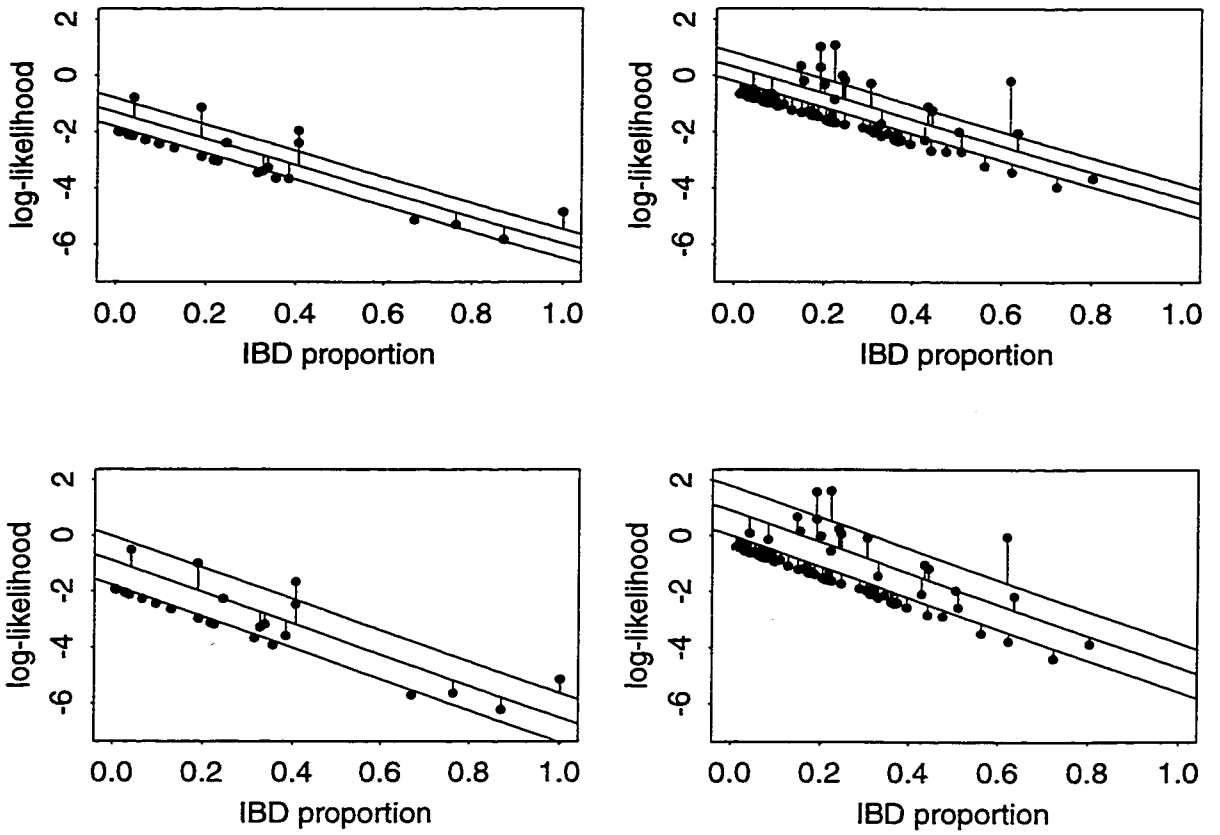

Figure D.3: **Markov approximation.** The fitted lines and deviation of observed log likelihood values from those lines for the Markov model. The upper row of plots show log likelihood ratios for the half-first-cousins-once-removed relationship, and the lower row of plots show log likelihood ratios for the second cousins relationship. The chromosomes are separated into two plots for each relationship, the left for chromosomes that are IBD at one end and the right for chromosomes non-IBD at both ends, to aid readability. In each plot the three lines are for chromosomes with, from bottom to top, 1, 2 and 3 IBD regions.

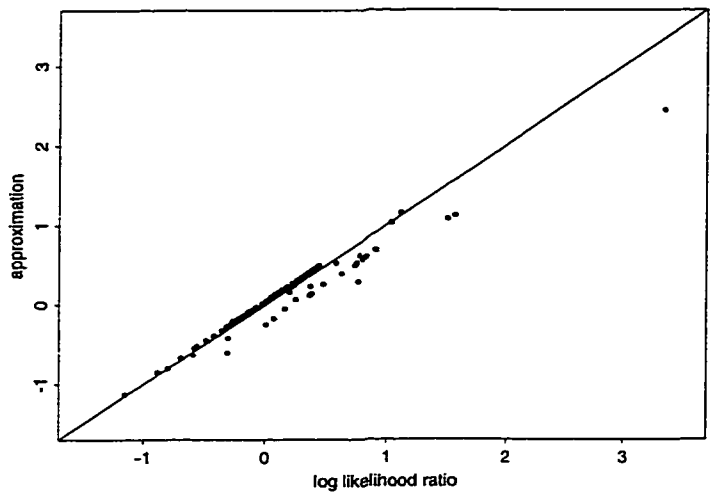

Figure D.4: The Markov approximation to the log likelihood ratio statistic is twice the difference between the log likelihoods for the second-cousins and half-first-cousins-once-removed relationships. Each point on the plot represents one of the 200 chromosomes of data. The points that are closer to the line (those well approximated by the model) tend to represent the chromosomes with only zero or one IBD regions.

ratio does not depend on the piece sizes.

When the data (of total length l) is cut into l/x pieces of length x we can count the number of pieces Y_1 that are non-IBD, the number Y_2 that are IBD, and the number Y_3 that contain IBD/non-IBD switches. The numbers (Y_1, Y_2, Y_3) are distributed trinomially with probabilities (p_1, p_2, p_3), where

$$p_1 \approx P(\text{non-IBD at 0 and } x) \approx 1 - k_1 - k_2 x$$

and similarly $p_2 \approx k_1 - k_2 x$ and $p_3 \approx 2k_2 x$ for small x.

As the pieces get smaller, the proportion of IBD pieces approximates the IBD proportion, p, and the number of pieces with switches equals the number of switches, s, in the data. The log likelihood is

$$\begin{aligned} l(k_1, k_2) &= y_1 \log(p_1) + y_2 \log(p_2) + y_3 \log(p_3) \\ &\approx \frac{1-p}{x} \log(1 - k_1) + \frac{p}{x} log(k_1) + s \log(2k_2 x) \end{aligned}$$

for small x.

When comparing two relationships with the same coefficient of kinship, k_1, but different two locus coefficients $K_2(x) \approx k_1 - k_2 x$ and $K_2'(x) \approx k_1 - k_2' x$ the log likelihood ratio $2(l(k_1, k_2) - l(k_1, k_2'))$ converges, as $x \to 0$, to

$$2\left[\frac{1-p}{1-k_1}(k_2' - k_2) + \frac{p}{k_1}(k_2' - k_2) + s\log(k_2/k_2')\right],$$

which is the same as the log likelihood ratio under the Markov model.

D.2 Renewal approximation

As a generalization of the Markov model, let us now assume that a renewal occurs at the start of each IBD and non-IBD region. That is we assume that the past history of the process becomes irrelevant at the start of each successive region. This assumption is valid for relationships in which one individual, or an ancestor of that individual, is the half-sib of the other individual, or an ancestor of that individual (the half-sib-type relationships described by Donnelly [7]), but is not generally hold. In particular the assumption is valid for the half-first-cousins-once-removed (which is half-sib-type) but does not hold for the second-cousins relationship. Even for relationships for which it is not correct, the assumption will be used to form an approximation.

D.2.1 Modeling region length distributions

We need to choose distributions for the lengths of the IBD and non-IBD regions. In the case of the greatgrandparent relationship and similar relationships, IBD regions have exactly an exponential distribution. Lengths of IBD regions for the cousins relationship are not exponentially distributed, but an exponential distribution seems to give a reasonable approximation. We will model IBD lengths as exponential with rate μ, as in section D.1.

IBD regions tend to occur in clumps, so that there are some short non-IBD lengths and some longer lengths. When the IBD process leaves an IBD state and enters a non-

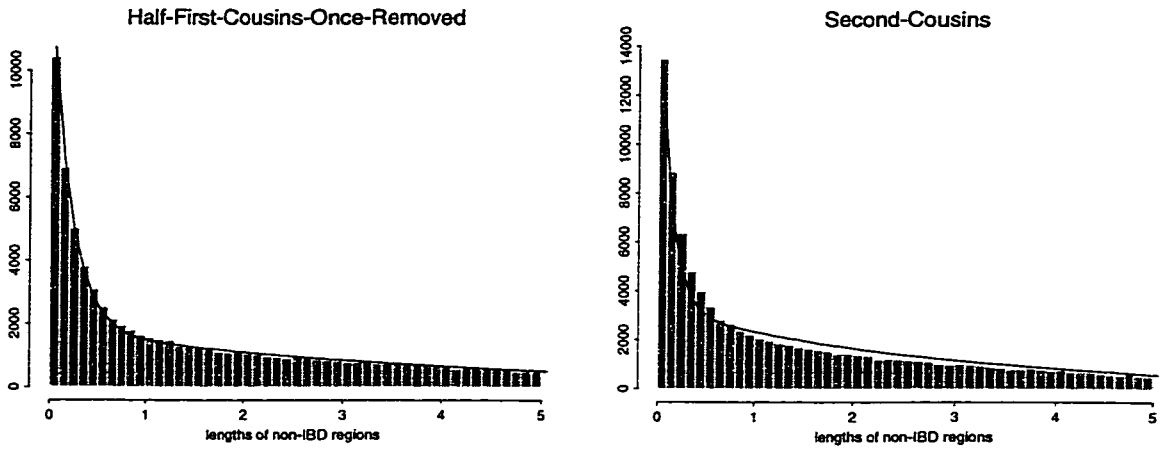

Figure D.5: **Non-IBD region lengths.** Histogram bars give the observed distribution of non-IBD region lengths from a simulation study. The overlaid line gives the modeled length distribution described above. Parameters for the model are $\nu = 5$, $\mu = 5$ and $q = 0.28$ for the half-first-cousins-once-removed relationship and $\nu = 8$, $\mu = 6$ and $q = 0.1875$ for the second-cousins relationship.

IBD state it can step back into the same, or a nearby, IBD state almost immediately. If it does not quickly return to IBD, the process tends to move further from the IBD states and becomes 'lost' deep in non-IBD territory and essentially memoryless. If we allow only one step for the IBD process to return to IBD before it is considered 'lost', the short non-IBD sections are exponentially distributed with rate ν, where ν is the number of meioses in the pedigree. The probability q of such a quick return can be found be investigating all possible two step moves from each IBD state. A long non-IBD section can be modeled as the sum of a short piece (exponential with rate ν) plus a long piece that is exponentially distributed with some rate λ, since the process becomes essentially 'lost' and memoryless after the first step. We will choose λ to be the value that makes the expected IBD proportion under the model to be k_1. This model has similarities with Aldous [1] Poisson clumping heuristic.

Figure D.5 shows the fit of this model for non-IBD lengths. The fit is reasonable, particularly for lengths from the half-first-cousins-once-removed relationship. This

model tends to give much better estimates of likelihood than the Markov model in section D.1, but its estimates of the log likelihood ratio statistic for second-cousins versus half-first-cousins-once-removed, shown in figure D.6, are generally worse than those from the Markov model.

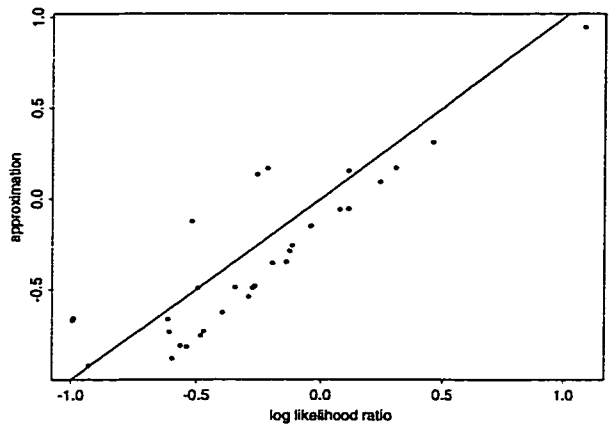

Figure D.6: Modeled renewal approximation to the log likelihood ratio for second-cousins versus half-first-cousins-once-removed with line of equality shown.

D.2.2 Empirical length distributions

Another approach to the problem is to use empirical length distributions, while still assuming renewals at the start of each IBD and non-IBD region. One way to do this would be to simulate a large amount of data and compile an empirical length distribution which would then be used in place of the modeled distributions above. Equivalently we can return to the Monte Carlo method of Browning. With the renewal assumption, we can estimate the likelihood for each region independently and multiply the estimates together at the end of the procedure instead of at each Monte Carlo iteration. This results in much faster convergence of the likelihood estimates than the non-renewal Monte Carlo procedure.

The estimates given by this renewal Monte Carlo method are only approxima-

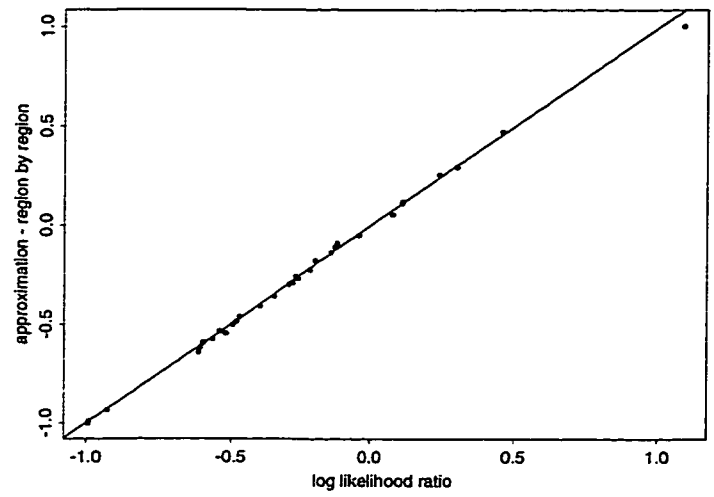

Figure D.7: Empirical renewal approximation to the log likelihood ratio for second-cousins versus half-first-cousins-once-removed with line of equality shown.

tions to the likelihood in cases such as second-cousins where the renewal assumption is not valid. Figure D.7 shows the approximation to the log likelihood ratio for second-cousins versus half-first-cousins-once-removed. The figure shows that the approximation is very good in this case.

D.3 Conclusion

The approximations presented here vary in quality of approximation to the relationship likelihood. The regression approach worked well for chromosomes with small numbers of regions, and the empirical renewal method did well with all the simulated data. The methods that didn't work were those that tried to fit an artificial distribution to the region lengths — the Markov and modeled length renewal approximations.

The IBD process is in general too complex to fit a simple model to, yet the information it contains is quite local. These local properties allow us to find simple ways to calculate an approximation to the likelihood, avoiding the huge amounts of

computation time involved in the full Monte Carlo likelihood approach.

The success of the empirical renewal method shows that, at least for the relationships considered here, while region lengths have a complicated distribution, the lengths are almost independent from region to region. In that sense the IBD process has a short memory.

VITA

Sharon Ruth Browning was born to Laurie and Shona Guy in Auckland, New Zealand on May 14th, 1973. In 1983 the Guy family moved to Papua New Guinea where Sharon and her sisters undertook their schooling by correspondence from the New Zealand Correspondence School. The family returned to New Zealand at the end of 1989 and in 1991 Sharon entered the University of Auckland. In 1995 Sharon received a Bachelor of Science degree with first-class honours in Mathematics from the University of Auckland. After teaching at the University of Auckland for six months in early 1995, Sharon moved to Seattle, Washington and entered the graduate program in Statistics at the University of Washington in Fall 1995. Sharon completed her Ph.D. in Statistics in Summer 1999.