UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Modélisation des paramètres de pénétrance incomplète et de phénocopie d'une méthode de cartographie fine d'une maladie complexe

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

SARAH VAHEY

FÉVRIER 2008

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Modeling Incomplete Penetrance and Phenocopy for Fine Mapping of Complex Diseases

THESIS

PRESENTED AS

PARTIAL REQUEST

FOR THE MASTERS IN MATHEMATICS

BY

SARAH VAHEY

FEBRUARY 2008

ACKNOWLEDGEMENTS

Before starting, I would like to thank the people who have helped and guided me throughout my time as a masters student at UQÀM.

Firstly, I would like to thank Fabrice Larribe, my thesis supervisor. I am grateful to have had the opportunity to be guided through my research by somebody so motivating, encouraging and also, someone who was always available to answer questions and help out.

Thank you to the professors of Statistics at UQÀM, and also the administration staff, in particular Manon, who have ensured that my learning experience during my masters went smoothly and was an enjoyable one.

Finally, thanks to Francis, my family and my friends for being there...

CONTENTS

LIST OF FIGURES vi

LIST OF TABLES viii

RÉSUMÉ ix

ABSTRACT x

INTRODUCTION 1

CHAPTER I
AN INTRODUCTION TO GENETICS 3

1.1 The Human Genome 3

1.1.1 Overview 3

1.1.2 DNA and Bases 4

1.1.3 Polymorphisms and Mutations 4

1.1.4 The Genotype-Phenotype relation 5

1.1.5 Binary and Quantitative Traits 6

1.2 Gene Association and Fine Mapping 7

1.2.1 Gene Inheritance 7

1.2.2 From Association to Fine Mapping 9

1.2.3 Graphical Representation of the Genealogy 12

1.3 Coalescent Theory and the Ancestral Recombination Graph 13

1.3.1 Introduction 13

1.3.2 Fundamental Insights in Coalescent Theory 13

1.3.3 The Wright-Fisher Model 15

1.3.4 Coalescence under the Wright-Fisher Model 16

1.3.5 The Ancestral Recombination Graph 17

1.4 Penetrance and Phenocopy 19

CHAPTER II
MAPARG AND OTHER FINE MAPPING METHODS 21

2.1 Simulation Techniques 22

2.1.1 Basic Monte Carlo Integration 23

2.1.2 Importance Sampling 24

2.1.3 Markov Chain Monte Carlo (MCMC) 25

2.2 MapArg 26

2.2.1 The Model 27

2.2.2 The Probability of a Mutation, Coalescent or Recombination Event 29

2.2.3 Recurrence Probability for the distribution of Q() 32

2.2.4 Importance Sampling Within The MapArg Framework 33

2.2.5 Composite Likelihood 35

2.3 Other Fine Mapping Methods 36

2.3.1 Decay of Haplotype Sharing 37

2.3.2 Fine Mapping Via The Shattered Coalescent 39

2.3.3 Tree LD 40

CHAPTER III
ACCOUNTING FOR INCOMPLETE PENETRANCE AND PHENOCOPY . . 44

3.1 Effects of Penetrance and Phenocopy 44

3.1.1 Introduction 44

3.1.2 The Structure of the Data 44

3.1.3 Simulating data with varying levels of Penetrance and Phenocopy . 46

3.2 Correcting For Penetrance And Phenocopy 50

3.3 Evaluating $P(H_0|H_{-1})$ 53

3.3.1 Method 1 53

3.3.2 Method 2 54

3.4 Simulations 57

3.5 Results In MapArg Accounting For The Penetrance Paramaters 59

3.5.1 A Description Of The Graphs in MapArg 59

3.5.2 Results On Simulated Data 60

3.5.3 Effect of sample size 66
3.5.4 Effect of half-window length, (l) 67
3.5.5 Results with the Cystic Fibrosis Data Set 69
3.6 Further Developments 71
3.6.1 Method 1 with Diploids 73
3.6.2 The Frequency of the Mutation for Diploid Data 75
3.6.3 Adaptations of other researchers' methods for MapArg 77
CONCLUSION 80
BIBLIOGRAPHY 82

LIST OF FIGURES

1.1 An illustration of the transmission of genes within a family. Males are represented by squares and females by circles. (from Larribe, 2003) . . 8

1.2 An illustration of the evolution of a segment of chromosomes over time. At $t = t_0$ a mutation occurs on one of the chromosomes and the sharing of material on this chromosome over several generations is recorded. The further forward in time, the smaller the amount of shared ancestral material around this mutation is. This is due to recombination (from Larribe, 2003). 11

1.3 An illustration of a genealogical tree in two different forms (from Larribe, 2003). 12

1.4 Realization of a coalescent process for 6 sequences on 10 generations (from Larribe, 2003) 14

1.5 An example of the ARG with four markers. Shaded-in boxes represent ancestral wild type alleles, un-shaded boxes represent non-ancestral material and half-shaded boxes, an ancestral mutant allele. (from Larribe, 2003) 18

2.1 Examples of possible coalescent events between different sequences. Shaded-in boxes represent ancestral wild type alleles, un-shaded boxes represent non-ancestral material and half-shaded boxes an ancestral mutant allele (from Larribe, 2003). 29

2.2 An illustration of window lengths for a sequence of 5 markers, where d equals the window length (from Larribe, 2007). 35

LIST OF TABLES

3.1 An example of the data MapArg works with. The first column represents the sequence in haplotype form, and the other two columns contain the frequency of cases and controls for each haplotype. 45

3.2 A data set with 5 sequences genotyped at 4 loci, *i.e.* 4 SNPs per sequence. There are 20 cases and 40 controls in total. 46

3.3 An example of the output data given from SimPenPhen when $f_1 = 0.85$ and $f_1 = 0.17$. There are 24 mutant and 36 non-mutant sequences in total. 47

RÉSUMÉ

Les méthodes de cartographie fine sont des modèles qui estiment la position d'un allèle mutant pouvant causer une maladie dans un groupe d'individus. Le travail de Larribe *et al.* (2002, 2003), MapArg, n'a pas tenu compte des paramètres de pénétrance jusqu'à maintenant. Ce mémoire démontre les effets de ces paramètres, soit la pénétrance et la phénocopie, sur la performance de MapArg, dans des populations haploïdes. De plus, deux méthodes que nous avons développées seront ensuite incorporées à MapArg dans le but d'améliorer son efficacité si il y a pénétrance et/ou phénocopie.

Les résultats démontrent que la phénocopie peut avoir une influence négative sur l'efficacité de MapArg. La pénétrance ne semble pas avoir d'effet majeur sur MapArg. La première méthode développée est un modèle simple qui n'apporte pas d'amélioration majeure de MapArg par rapport à ce même modèle sans ajustement. Par contre, cela procure un point de départ pour les développements futurs dans les populations diploïdes. La deuxieme méthode améliore l'efficacité de MapArg sous certaines conditions, en particulier, si la taille de l'échantillon est assez grande. La deuxieme méthode fonctionne également très bien pour les données réelles de la Fibrose Kystique (Kerem *et al.*, 1989).

Mots clés: phénocopie, pénétrance, pénétrance incomplète, cartographie fine

ABSTRACT

Fine mapping methods are models that provide an estimate for locating a mutation causing a given disease among a group of individuals. MapArg, the work of Larribe *et al.* (2002, 2003), did not take penetrance parameters into account to date. This thesis shows the effect of these parameters, namely penetrance and phenocopy, on the performance of MapArg for haploid populations. Also, two different methods are developed and incorporated into the MapArg framework with the goal of increasing efficacy of MapArg in the presence of penetrance and/or phenocopy.

Results show that phenocopy can strongly effect MapArg's efficiency while penetrance does not have much of an effect. The first Method developed is a simple model that does not prove much more efficient than MapArg without any adjustment; however, it provides the groundwork for further development when diploid populations will be modeled. Method 2 has shown to improve the efficiency of MapArg under certain conditions, in particular, when the sample size is large. This method also greatly improves the performance of MapArg with the Cystic fibrosis data (Kerem *et al.*, 1989).

Keywords: phenocopy, penetrance, incomplete penetrance, fine mapping

INTRODUCTION

Gene mapping of complex diseases is an ongoing research in the field of genetics. One of the primary goals of this research is to locate the position of a mutation(s) (or causal gene(s)), that causes a given disease. Fine mapping, a branch of gene mapping, uses information from previous studies that have ascertained an approximate location for the mutation(s) or causal gene(s) and concentrates on pinpointing the exact location. Larribe *et al.* (2002) and Larribe (2003) have developed a fine mapping method called MapArg that estimates the position along a chromosome of a mutation responsible for a given disease. Approximating complex biological processes by means of mathematical models is a difficult procedure and usually some hypotheses that are not always realistic are necessary in order for these models to be feasible. The research of Larribe, and the fine mapping methods of his contemporaries, are constantly evolving over time, incorporating models for biological aspects that were not previously accounted for.

One assumption that MapArg has worked with to date is that the disease being studied has complete penetrance and no phenocopy. In biological terms this means the following: if individuals are affected by a certain disease they automatically carry the mutation causing this disease and likewise, if individuals are not affected by disease they do not carry the mutation. It is well known however, that for complex diseases, there exists incomplete penetrance and phenocopy, e.g. Breast Cancer: some women suffer from breast cancer without carrying the causal gene (phenocopy) and other women carry the causal gene but do not suffer from breast cancer (incomplete penetrance). These phenomenon collectively known as the penetrance parameters are currently being taken into account either directly or indirectly, by McPeek and Strahs (1999), Morris *et al.* (2002) and Zöllner and Pritchard (2005) in different ways.

The goal of this thesis is to study the effect of incomplete penetrance and phenocopy on the performance of MapArg, and also to develop some models that can account for these parameters within the MapArg framework. This body of work is composed of three chapters. Chapter 1 contains an introduction to the biological notions that are necessary to understand fine mapping. The mathematical models upon which MapArg and other fine mapping methods are based are also presented in Chapter 1. A detailed explanation of MapArg is discussed in Chapter 2 along with a review of the fine mapping methods in the literature. Particular attention is given to the way in which other research methods have taken incomplete penetrance and phenocopy into account. The third and final chapter and is in fact the crux of the thesis and the original work contained here encapsulates the main goal of this thesis. The effects of the penetrance parameters on MapArg are shown. Following this are two methods that have been developed to take these parameters into account within the MapArg framework. Results of the performance of each method are then presented and discussed. Also ideas for future development is discussed.

CHAPTER I

AN INTRODUCTION TO GENETICS

1.1 The Human Genome

1.1.1 Overview

Although the subject at hand is to develop existing statistical methods for analyzing data, we are nonetheless working in the domain of Population Genetics. For this reason, it is necessary to introduce some basic biological notions. This should facilitate the reader in understanding the motivation behind our research and also the methods that will be presented in this thesis.

Within each nucleus of each human cell, there are 23 pairs of chromosomes, 22 paired autosomes along with two sex chromosomes. These 23 pairs of chromosomes contain all of the genetic information for this individual. Each chromosome is formed from a long piece of DNA (deoxyribonucleic acid), a coiled double-stranded molecule that winds itself up inside the chromosome. About 2 % of the DNA in a chromosome represents genes. Genes are what determine a person's characteristics and there are between 30 000 and 40 000 of them in the nucleus. This complete set is called the human genome.

Each chromosome carries a couple of thousand genes and many of these are common to all human beings. In fact approximately 99.9 % of all humans DNA is identical. It is the other 0.1 % that distinguishes us from one another, an example of a characteristic or trait is eye color. An important characteristic that is of interest to us throughout this

thesis is a person's susceptibility to a disease. Environmental factors, such as lifestyle (e.g. smoking and nutrition), also influence our susceptibility to disease. Modeling the relationship between a person's genetic makeup and environmental factors is necessary to understand how an individual has contracted an illness. If a person is diagnosed with a particular illness, and we have information on their lifestyle and environment as well as their genetic makeup, it would be very informative and could assist in finding the location of the gene(s) responsible for this illness.

1.1.2 DNA and Bases

The DNA molecule is a double helix. It has two strands with links between them. Each link between the strands is made from a pair of nucleotide bases. There are four types of bases called Adenine(A), Guanine(G), Cytosine(C) and Thymine(T). Each base pairs up with its complementary base:

- A pairs with T
- G pairs with C

The sequence of these base pairs is unique to any individual apart from identical twins, who have the same DNA.

The DNA in our chromosomes has 3 000 000 000 base pairs (or 3000 megabases), noted 3000Mb. A single gene is represented by a few thousand bases, so given that there are between 30 000 and 40 000 genes, around 150Mb carry useful information. Thus approximately 98 % of our DNA is data that does not carry any genetic information.

1.1.3 Polymorphisms and Mutations

Each time a cell divides, the chromosomes in the new cell carry a copy of the original chromosomes. A polymorphism occurs when there is an error made during the DNA replication. Sequence variations result in different forms of the same gene. These

forms of the same gene are called alleles. Polymorphisms are common differences in the sequence of DNA, occurring in at least 1% of the population. By definition, Single Nucleotide Polymorphisms (SNPs) are the smallest possible change in DNA. Alleles that occur most frequently in the population are referred to as wild type alleles. For example, if there are two possible alleles at a locus called A and a, and A is the more common of the two then it is called the wild type allele for that particular locus.

Mutations are less common differences in DNA sequences, occurring in less than 1% of the population. They can be passed down from a parent to a child, can occur during conception or may even be acquired during an individuals lifetime. A mutation can arise in response to an environmental factor such as exposure to toxins or diet. A large proportion of DNA variation has no effect but some changes within genes can contribute to the susceptibility of a disease.

1.1.4 The Genotype-Phenotype relation

Each gene or pair of alleles has a particular position along the chromosome. The locus is the term used for the position of a gene but it may also refer to the position of a marker. A marker is a general term for a trait or DNA segment that is easily identified. Markers are useful as they may be closely linked to a gene or trait that is difficult to identify. In the work presented here we will be using SNPs, whose positions are well known, as markers along the chromosome. If we look at a set of loci along a chromosome the collective name for the pairs of alleles found at each locus is the haplotype.

The genetic material transmitted from one generation of humans to the next is done so in pairs of chromosomes. Therefore, when studying the ancestry of a family, we should simultaneously account for the transmission of haplotype on 2 chromosomes, one from the mother and one from the father. Humans are said to be diploid for this very fact, but we shall see in subsequent chapters that modeling diploid populations adds a level of complexity to an analysis. Normally, to simplify matters, humans are regarded as haploids: instead of looking at pairs of haplotypes, individual chromosomes are studied

independent of their pair. In other words, we view the genetic makeup of a human 46 separate sequences as opposed to 23 pairs of chromosomes.

The genotype represents the allelic composition of an organism that determines a trait, whiles the phenotype is the physical expression of this allelic composition *i.e.* the physical trait. Genotypes and phenotypes are very closely linked: a combination of alleles on one or several chromosomes (the genotype) manifests itself physically as a trait or characteristic (the phenotype). Unfortunately the relationship between genotype and phenotype can be complicated by environmental factors. The phenotype may also change over time due to aging or environmental changes. Further complexity may arise due to gene interaction, when a physical trait does not express itself unless a particular combination of more than one gene is present.

A dominant relationship may exist between alleles at the same locus. A dominant allele is one which automatically influences the phenotype regardless of the other allele type at the locus. If an allele is not dominant then it is known as a recessive allele. The only way in which a recessive allele can be expressed physically is if an individual carries two recessive alleles at the same time.

1.1.5 Binary and Quantitative Traits

The phenotype of interest may have a binary or quantitative trait. Binary traits are ones that have two possible outcomes only: either the characteristic of interest is present or not present. It is conventional to say the status of a binary trait phenotype is either *affected* or *unaffected*. An example of a phenotype with a binary trait is Cystic Fibrosis. In this case, if a collection of individuals is tested for the illness, they will be grouped into two sets, those who have Cystic Fibrosis (affected) and those who don't (unaffected). In statistical terms, affected individuals are considered as cases and unaffected individuals as controls. Quantitative traits are those which are measured on a continuous scale. Blood pressure is an example of one such trait. We will be working with binary trait phenotypes only in our work and henceforth, will refer to individuals as being either

cases or *controls*.

In gene-mapping, a series of sequences are analyzed to see if it is possible to find the location of a gene responsible for a particular trait. Usually the trait we are interested in is a disease or illness. We call the gene responsible for this trait, a trait influencing mutation (TIM). Assuming a relationship between genotype and phenotype without any sort of distortion would lead us to conclude the following: cases possess the TIM whiles controls do not possess the TIM. Unfortunately, it is rare that a distortion free relationship between genotype and phenotype exists. Moreover, as humans are diploid, they have 0, 1 or 2 copies of the TIM; if the model is dominant, people are cases with 1 or 2 mutations, whereas in a recessive model, if people are controls, they can have 0 or 1 TIM. So the relation between the TIM and the phenotype is far from direct. We will make some simple assumptions regarding diploidicity later, in order to simplify development. In Chapter 3, we will introduce a method that models this complex relationship within the context of gene mapping.

1.2 Gene Association and Fine Mapping

1.2.1 Gene Inheritance

Genes are passed down from parent to child during a process known as meioses. Figure 1.1 shows alleles at three loci being transmitted from two parents to seven children. Individuals represented by a square are males and those represented by a circle are females. We can see that children 3, 4, 7, 8 and 9 receive complete blocks of genetic material from the paternal segment. However, children 5 and 6 receive alleles from both chromosomes of the father. This is because DNA segments of the maternal chromosomes have switched during meiosis and thus these children receive material from the paternal grandmother and paternal grandfather at the same time. This process of crossing over of segments of a parental sequence is known as recombination.

The probability that there is a recombination between two loci during meioses is known

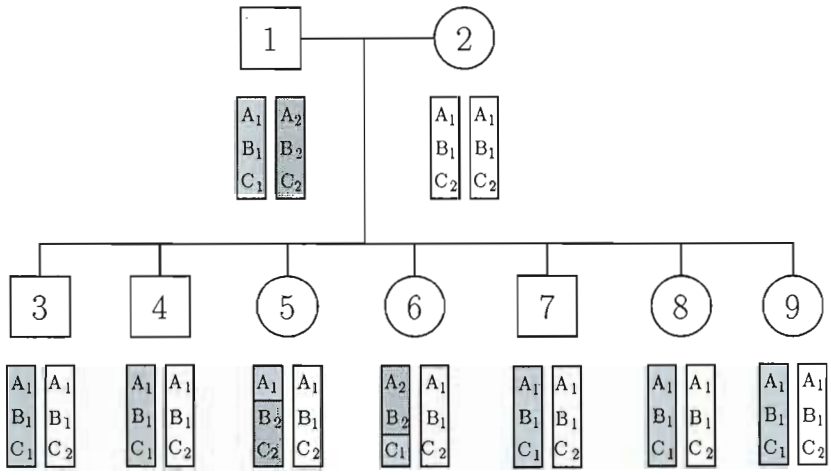

Figure 1.1 An illustration of the transmission of genes within a family. Males are represented by squares and females by circles. (from Larribe, 2003)

as the recombination fraction, denoted by θ. The rate of recombination is a function of the distance between two alleles. When two loci are on different chromosomes they are not linked as they are transmitted independently of one another, hence the recombination fraction $\theta = \frac{1}{2}$. This is because either allele has an equal probability of being transmitted to an offspring. Loci that are on the same chromosome have a chance of recombination. In this case $\theta < \frac{1}{2}$ and the alleles at these loci are considered *linked.* The further apart the two loci are positioned, the closer θ is to $\frac{1}{2}$ and the weaker the linkage is between the alleles. This fact leads to the concept of genetic distance. Although related to physical distance there is no simple function that relates them. The reason for this is that the rate of recombination is not uniform across the chromosomes. Moreover, the recombination rate is different for males and females. We work with genetic distances and the unit of measurement for this distance is the Morgan. The Morgan is defined as the unit of distance where exactly one recombination event is expected to occur (Olson *et al.* 1999).

1.2.2 From Association to Fine Mapping

Association analysis is widely used to infer an approximate location of a TIM as follows: a segment of DNA is genotyped at well known positions called markers (see Section 1.1.4) for as many relations as possible within a family. If enough information is available the inheritance pattern for this pedigree can then be traced back in time. This pattern is then compared to the inheritance pattern of the phenotypes observed. Regions where the two patterns are highly correlated are thought to be linked and so an approximate location of the TIM is estimated.

Consider two loci where one is a TIM locus with alleles A and a occurring at relative frequencies p_A and p_a and the other is a marker locus with alleles B and b occurring at relative frequencies q_B and q_b. There are a total of four possible haplotypes, namely AB, Ab, aB and ab, with corresponding relative frequencies h_{AB}, h_{Ab}, h_{aB} and h_{ab}. If the two loci are not linked, that is to say independent of one another, we can expect that $h_{AB} = h_{Ab} = h_{aB} = h_{ab}$.

Denote h_{AB0} as the relative frequency of the haplotype AB at the present generation. As previously defined, the recombination fraction is θ, and $(1 - \theta)$ is the probability of being a non-recombinant in the next generation. Assuming independence of loci, if a haplotype is non-recombinant in the next generation, then the probability of the haplotype being AB again is h_{AB0}. On the other hand, if there is a recombination event, the probability of the haplotype being AB in the next generation is the product of the two allele frequencies $p_A q_B$ by the hypothesis of independence. Therefore the probability that the haplotype AB is transmitted to the next generation is equal to

$$h_{AB1} = (1 - \theta)h_{AB0} + \theta p_A p_B, \tag{1.1}$$

and the change in the frequency of haplotype AB from one generation to the next is given by

$$h_{AB1} - h_{AB0} = \theta(p_A p_B - h_{AB0}). \tag{1.2}$$

As we know, if there is no association between alleles from different loci then $p_A p_B =$

h_{AB0} and the haplotype frequency does not change between generations. If there is allelic association the haplotype frequencies will change from generation to generation and the change is in proportion to θ. This non-random association of haplotypes is referred to as Linkage Disequilibrium (LD). Over time a population with linkage disequilibrium will approach equilibrium. The number of generations this will take is a function of θ, the recombination fraction. If (1.1) is written as

$$h_{AB1} - p_A p_B = (1 - \theta)(h_{AB0} - p_A p_B), \tag{1.3}$$

it can be seen that the distance between the relative frequency of h_{AB} and its equilibrium value $p_A p_B$ decreases by a factor of $(1 - \theta)$ at each generation. So if θ is small, the breakdown of LD between the TIM and marker will take many more generations than if θ is large.

In fine mapping studies, researchers make use of the fact that there is linkage disequilibrium between a TIM and alleles in close proximity to it. Even though the size of the segment around the mutation decreases over time due to recombination, there is less likelihood of recombination between loci that are positioned closely together on the chromosome. Fine mapping as the name suggests, looks at a small area with markers that are thought to be tightly linked to the TIM, using previous information from association studies.

Figure 1.2 describes the relation of a group of sequences, some of which have been effected by a TIM. At time t_0 a mutation is produced on one of the sequences. After several generations the mutation has been transmitted to offspring during meioses (t_1). Notice that most of the chromosomal region about the TIM remains intact, but already some of the original genetic material has been lost due to recombination. By the time t_2 is reached, there is only a small area around the TIM that is similar to all sequences effected by it. If we just look at the segment of the sequences between the two dashed lines it is clear that those effected by the TIM are more similar to each other than to the population at large.

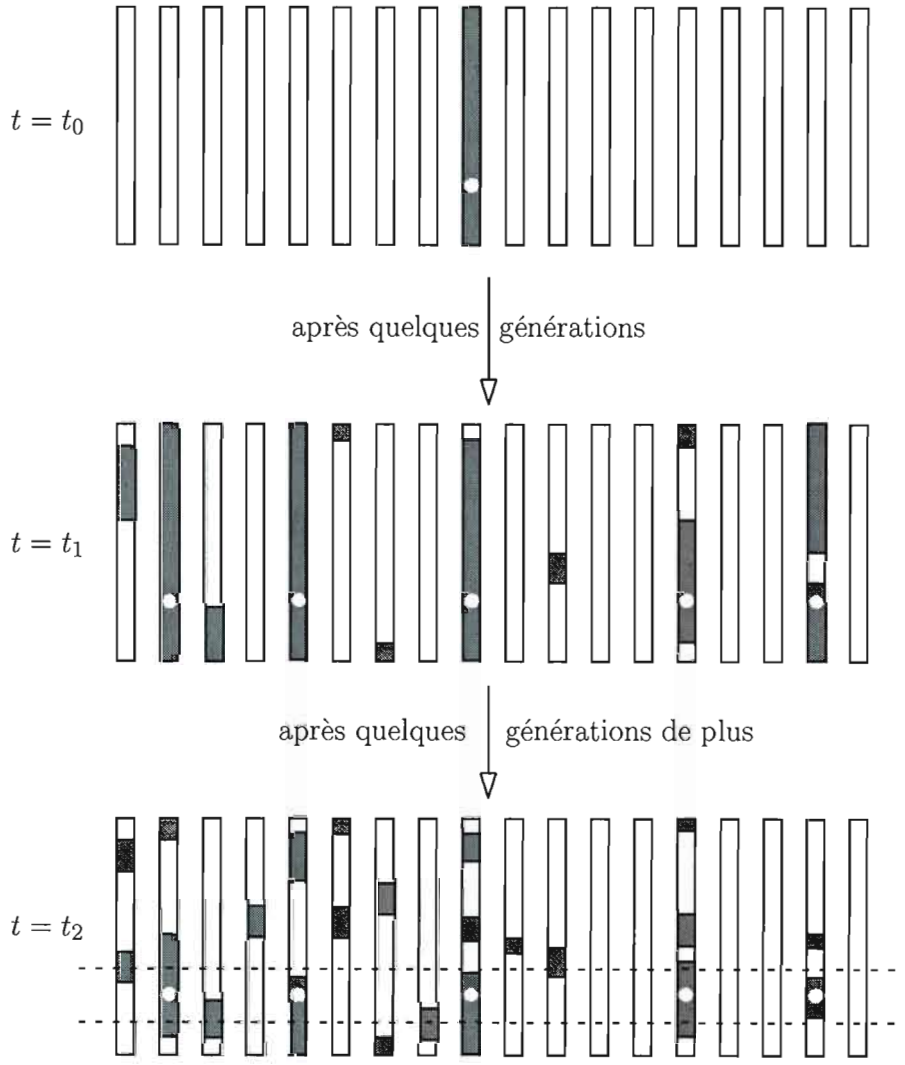

Figure 1.2 An illustration of the evolution of a segment of chromosomes over time. At $t = t_0$ a mutation occurs on one of the chromosomes and the sharing of material on this chromosome over several generations is recorded. The further forward in time, the smaller the amount of shared ancestral material around this mutation is. This is due to recombination (from Larribe, 2003).

1.2.3 Graphical Representation of the Genealogy

The recording of the ancestry of a group of individuals, that may or may not be related to one another, is called a genealogy. It is important to note that although these individuals may be unrelated, they still share common alleles, a fact that allows us to construct their genealogy. As the graphical representation of a genealogy resembles a tree, we will often refer to the genealogy of a sample as the sample tree. The two most important features of the genealogy, for the purposes of fine mapping are the topology (shape) and the branch lengths. Figure 1.3 below is an example of a genealogical tree.

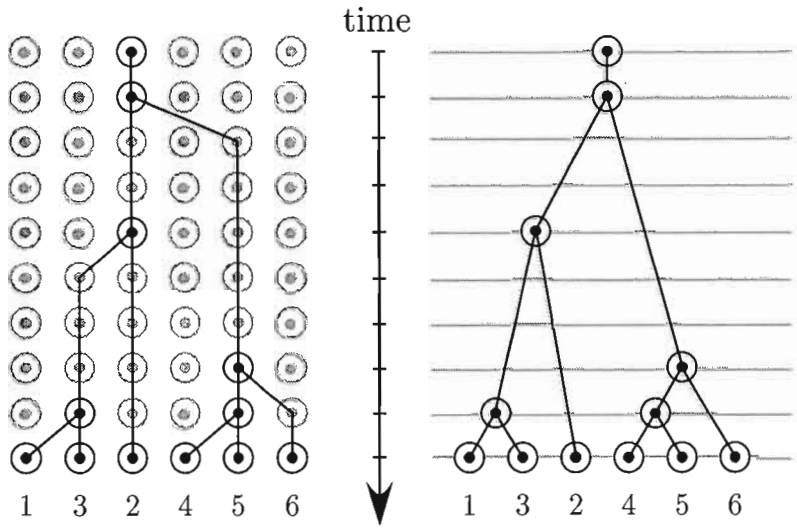

Figure 1.3 An illustration of a genealogical tree in two different forms (from Larribe, 2003).

The sequences appearing at the bottom of the tree are often referred to as the leaves of the tree, and represent the sample history at the present time or time zero (τ_0). When all of the sequences join together at the top of the tree, this time is denoted as τ_*. The apex of the tree is referred to as the most recent common ancestor (MRCA), as indeed it represents the nearest time in history when all of the sequences have an ancestor in common. Two or more sequences, also referred to as lineages join together at a node

and the interval of time between each event or each node is recorded. As previously mentioned, it is assumed that sequences possessing the TIM have a common ancestry closer back in time than sequences not possessing the TIM have. By looking at the leaves of a tree, we are given a good indication as to which sequences have the TIM and which don't as the sequences that bunch together at the bottom of the tree (i.e the most similar sequences) will be expected to possess the TIM. Furthermore, events such as the occurrence of a mutation are recorded at each node, enabling us to trace the TIM from the point in time that is occurs to the present. The construction of the tree as well as the recording of mutation events has been in effect for several decades and is known as coalescent theory. This theory, primarily developed by Kingman (1982) will be explained in the following section.

1.3 Coalescent Theory and the Ancestral Recombination Graph

1.3.1 Introduction

The modeling of a genealogy by studying events in the ancestry of the sample is a stochastic process that is known as the coalescent. Kingman (1982) developed the mathematical model of coalescent theory and it has since proven a useful way to model data in population genetics. There are two fundamental insights that result in coalescent theory being an effective approach in modeling genetic data.

1.3.2 Fundamental Insights in Coalescent Theory

Insight I

Every sample of sequences have a genealogy, whether it displays variation or not. Under the assumption that all variation is selectively neutral, *i.e.* an individual genotype has no influence on the number of offspring it produces, the mutation process may be separated from the genealogy process. This means that a genealogy may be constructed for a sample, after which a mutation model my be superimposed. To quote Nordborg

(2001) " In classical terms 'state' may be separated from 'descent'".

Insight II

Starting off with a sample of n sequences, each sequence randomly chooses a parent out of n sequences in the previous generation. The genealogy of the sample can be constructed by tracing the coalescence of each sequence until there is one ancestor common to the sample, the MRCA. Thus the complete genealogy of a sample may be modeled without needing to know about the rest of the population. On the left hand side of figure 1.4 we see a sample of 6 sequences that randomly choose a parent in the previous generation while on the right, the ancestry of these 6 sequences is followed until the last generation. This sample shares a common ancestor in the second generation.

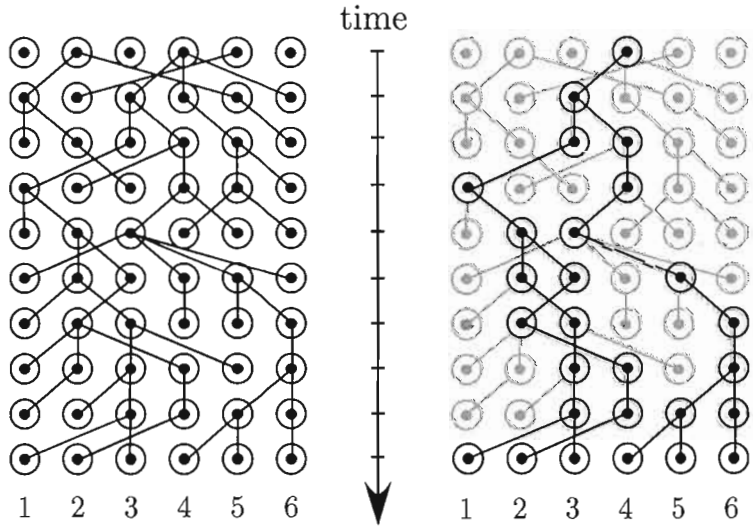

Figure 1.4 Realization of a coalescent process for 6 sequences on 10 generations (from Larribe, 2003)

It can easily be seen that given these insights, modeling a sample of sequences backwards in time with the coalescent may be more efficient than modeling forward in time, a process which requires looking at a whole population over a long period of time.

1.3.3 The Wright-Fisher Model

Constructing a model which aptly describes our population yet is mathematically tractable at the same time, is not an easy task. One such model which is widely used in statistics today is the Wright-Fisher Model, so called, as both Fisher and Wright used this model when dealing with genetics. The model makes the following assumptions:

- Generations are non-overlapping;
- The population size remains constant, *i.e.* no migration;
- The population is finite in size;
- There is no selection, *i.e.* an individuals genotype does not influence the probability of reproduction.

It is assumed ancestors of the present generation are obtained by random sampling with replacement of the previous generation. Consider two alleles A and a, at a particular locus. The population size is N, i alleles of which are A and $N-i$ of which are a. Allele frequencies for A and a among this population are $\frac{i}{N}$ and $\frac{(N-i)}{N}$ respectively.

The probability that allele A will have j copies in the next generation given that it has i copies in this generation happens to be modeled by the binomial distribution:

$$\binom{N}{j} p^j (1-p)^{N-j}, \qquad 0 \leq j \leq N,$$

where p is the frequency of allele A, *i.e.* $p = \frac{i}{N}$. Letting C_1 represent the number of copies of allele A in generation 1, the present generation being 0, then C_1 is a binomial random variable. Thus

$$E(C_1) = Np = i,$$
$$Var(C_1) = Np(1-p).$$

The frequency of A will drift going from past to present, either becoming extinct or reaching the population size N.

1.3.4 Coalescence under the Wright-Fisher Model

Constructing the genealogy of a sample consists of knowing the topology and branch lengths of the tree. Under the Wright-Fisher model and due to selective neutrality, all sequences are equally likely to coalesce.

Lets take a sample of n sequences. A sequences randomly chooses a parent from the previous generation independently of all other sequences. The probability of two sequences having the same parent is $\frac{1}{N}$ and the probability that two sequences have distinct parents is $1 - \frac{1}{N}$. As generations are non-overlapping, the probability of being distinct for t generations is $(1 - \frac{1}{N})^t$. If we re-scale the time so that one unit corresponds to N generations we have

$$P(\text{2 distinct lineages}) = \left(1 - \frac{1}{N}\right)^{N\tau}.$$

As N goes to infinity,

$$\left(1 - \frac{1}{N}\right)^{N\tau} \rightarrow \exp^{-\tau}.$$

Thus, the coalescent time for a pair of lineages is Exponential with mean 1. Considering k lineages, the probability of k distinct lineages in the previous generation is

$$\prod_{i=0}^{k-1} \frac{N-i}{N} = \prod_{i=0}^{k-1} \left(1 - \frac{i}{N}\right) = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right).$$

Now we want the probability of k distinct lineages for exactly t generations. This is the same thing as not having any coalescent event for t generations and then having a coalescence in the $t + 1th$ generation, which is

$$\left(1 - \frac{\binom{N}{2}}{N}\right)^t \left(\frac{\binom{N}{2}}{N}\right).$$

As N tends to infinity, by the same argument as above we have that the time to coalescence for k lineages is Exponential with mean $\frac{2}{k(k-1)}$.

Each step back in time the number of lineages decreases by one. Thus $T(k)$ is the time from k to $k-1$ lineages. As each coalescent event is independent, the $n-1$ coalescent times, $T(n)$, $T(n-1)$, ..., $T(2)$ are mutually independent exponential random variables.

Most of the variability in tree height is determined by $T(2)$. Intuitively this makes sense as the more lineages there are the more likely it is to have a coalescence. With only two sequences left to coalesce, this make take a very long time.

When using coalescent theory to estimate trees for a sample of sequences, thousands of trees are constructed. The idea is to find the trees that lead to a lot of information about the TIM position. Thus we are looking for informative trees. It is important to note that the coalescent is used to estimate genealogies and collectively use this information to find the TIM.

1.3.5 The Ancestral Recombination Graph

Griffiths and Marjoram (1996a, b) introduced the Ancestral Recombination Graph (ARG), an extension of the coalescent that takes recombination events into account. As a recombination can result in one sequence branching into two separate sequences in the past, we no longer deal with a tree, so a graph is now used. An example of an ARG is presented in figure 1.5.

The graph depicts three possible events that can occur when looking backwards in time. A coalescent event can occur resulting in two sequences joining together in a previous generation. Mutation events leave the number of lineages unchanged, a single marker changes however. One lineage branching into two lineages in the past represents a recombination event. Recombination results in both ancestral and non ancestral material appearing on the ARG. When a recombination event occurs, the sequence is split into two parental sequences, the left parental sequence and the right one. The left parental sequence shares the same genetic material as the "child" from the start of the sequence to the point of recombination. The right parental sequence shares the same genetic material as the child from the point of recombination to the end of the sequence.

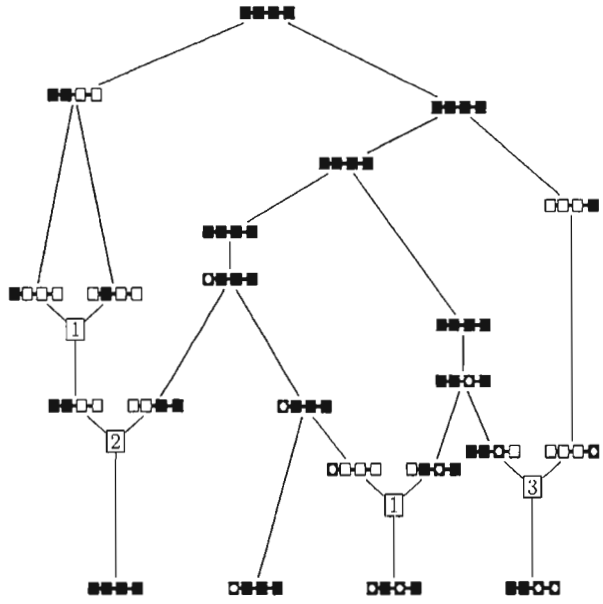

Figure 1.5 An example of the ARG with four markers. Shaded-in boxes represent ancestral wild type alleles, un-shaded boxes represent non-ancestral material and half-shaded boxes, an ancestral mutant allele. (from Larribe, 2003)

Recombination is a phenomenon that plays a large role in linkage disequilibrium, the non-random association of haplotypes, and consequently in fine mapping genetic sequences. Although taking recombination into account when modeling a sample genealogy can complicate computation, it does provide a more realistic representation of the behavior of genes.

1.4 Penetrance and Phenocopy

Thus far, the biological notions along with mathematical models involved in estimating the location of the TIM have been discussed. This last section of Chapter 1 will describe one final biological phenomenon that has a large effect on the ability of fine mapping methods to accurately locate the TIM, namely penetrance and phenocopy.

As already mentioned, the relationship between the genotype and the phenotype may be quite complex. If however, some additional information was available on the trait of interest, if may be possible to model the genotype-phenotype relationship with more precision. Let us introduce the following parameters:

$$\begin{aligned} \text{Penetrance} &= f_1 = Pr(\text{Affected}|TIM), \\ \text{Phenocopy} &= f_0 = Pr(\text{Affected}|\overline{TIM}), \end{aligned}$$

where $\overline{TIM}$ is the compliment of TIM. If $f_1 = 1$ and $f_0 = 0$ then we say there is complete penetrance and no phenocopy. When dealing with complex diseases however, it is usually the case that $f_1 < 1$ and $f_0 > 0$.

More generally, humans have genes by pair, so we have to define penetrance by three probabilities:

$$\begin{aligned} \tilde{f}_0 &= Pr(\text{Affected}|0 \text{ TIM}), \\ \tilde{f}_1 &= Pr(\text{Affected}|1 \text{ TIM}), \\ \tilde{f}_2 &= Pr(\text{Affected}|2 \text{ TIM}). \end{aligned}$$

If a disease is dominant, then $\tilde{f}_1 = \tilde{f}_2$, and if a disease is recessive $\tilde{f}_0 = \tilde{f}_1$. We expect

that $\tilde{f}_0 <= \tilde{f}_1 <= \tilde{f}_2$. However, as we will see later, we suppose for simplification of some parts of the present work that penetrance is just defined by f_1 and f_0.

Linkage disequilibrium (LD) is one of the fundamental insights upon which fine mapping methods are built. With LD we make use of the fact that sequences possessing a rare mutation (*i.e.* cases) are likely to share more ancestral material amongst one another than with the rest of the sample (*i.e.* controls). If incomplete penetrance exists among the sample of sequences being modeled, it is no longer clear which cases possess the TIM anymore. Likewise, if there are phenocopies among the sample it could turn out that there are cases who have contracted the disease being studied, without actually possessing the TIM. This could result in looking for ancestral linkage where it may not exist and could render fine mapping methods less accurate.

It is of great importance therefore, to try and estimate these parameters in the context of fine mapping. Chapter 3 will discuss the effect of various levels of penetrance and phenocopy on the fine mapping method MapArg. Also, some methods for modeling these parameters are developed. Here is a short example for illustrative purposes

An Example of the effect of penetrance and phenocopy

20 people are known to have the TIM for disease A and 40 people are known not to have this TIM. It is also known that the penetrance for this disease is .85 and the phenocopy level it .17. Thus if we want measure the number of people that are sick, a very good approximation is $20 * .85 + 40 * .17 = 23.8$ or 24 people.

CHAPTER II

MAPARG AND OTHER FINE MAPPING METHODS

In 2002, Larribe *et al.* developed a fine mapping method using likelihood theory and Importance Sampling simulation techniques. This body of work, henceforth called MapArg, produces a likelihood estimate of the location of the TIM. Up until now MapArg has been implemented with the assumption that there are no phenocopies and there is complete penetrance *i.e.*, every case has the TIM and no control has the TIM. As we know, most illnesses have some level of phenocopy and not necessarily complete penetrance. Breast Cancer for example may be contracted due to environmental factors, resulting in phenocopies and not every female that possesses a gene causing breast cancer will actually develop the disease and so there is incomplete penetrance.

Before we go on to describe our suggested methods for dealing with incomplete penetrance and phenocopy in fine mapping, we will describe MapArg in some detail. Other researchers that have currently developed fine mapping techniques for locating the TIM, have had to model incomplete penetrance and phenocopy also. Their methods will be described with an emphasis on how they deal with these parameters. In particular, the work of McPeek and Strahs (1999), Zöllner and Pritchard (2005) and Morris *et al.* (2002) will be discussed.

2.1 Simulation Techniques

Simulation can be used to approximate integrals when no exact solution can be found, which is the case in most gene mapping methods, as they involve multi-dimensional integration. There are two very popular methods of simulation in use today, Importance Sampling (IS) and Markov Chain Monte Carlo (MCMC). MapArg employs IS techniques while the authors mentioned above use MCMC methods to deal with their integration.

In all cases, an estimate for the position of the TIM is sought, whether by Maximum Likelihood or Bayesian Inference. There are several parameters to be estimated in order to estimate genealogies, but let us assume for simplicity that we are dealing with one parameter θ. If we denote $L(\theta)$ as the likelihood of the data in a sample *i.e.* $L(\theta) = P(\theta|D)$, then $\hat{\theta}$ is the estimate that maximizes this likelihood under maximum likelihood inference. In other words $\hat{\theta}$ is the maximum likelihood estimate (MLE) for θ. Bayesian inference incorporates a prior distribution about θ into the model and the distribution of θ is given as

$$P(\theta|D) = \frac{L(\theta)P(\theta)}{P(D)}.$$

A point estimate for θ is then reported and is usually the mode of the distribution $P(\theta|D)$, where D represents the data. The previous chapter introduced the idea of using a genealogical tree of a sample of unrelated individuals to find the position of a mutation (see Section 1.3). Estimating the tree, T, at each potential location of the TIM is pivotal in finding its true position. The likelihood of the data can be written in terms of the sample genealogical trees as follows:

$$L(\theta) = P(D|\theta) = \sum_{T} P(D|T,\theta)P(T|\theta).$$

Now $P(D|T,\theta)$ and $P(T|\theta)$ can be calculated more easily than $P(D|\theta)$ as we shall shortly illustrate. The quantities in the tree evaluation are continuous thus giving an

integral instead of a sum:

$$L(\theta) = P(D|\theta) = \int P(D|T,\theta)P(T|\theta)dT.$$

Calculating this integral is extremely complex and time consuming. Stephens (2001) notes that the number of tree topologies relating ten chromosomes is 2 571 912 000! This is where approximation techniques using simulation can be very useful.

2.1.1 Basic Monte Carlo Integration

IS and MCMC techniques build on basic Monte Carlo simulation so we shall describe the principles of this before explaining IS and MCMC in detail. Suppose we want to evaluate

$$I = \int_a^b h(x)dx,$$

where $h(x)$ is a very complicated function for which no closed form solution can be yielded. I can also be written as

$$I = \int_a^b h(x)dx = \int_a^b w(x)f(x)dx,$$

where $w(x) = h(x)(b-a)$ and $f(x) = \frac{1}{(b-a)}$. Now $f \sim Uniform(a,b)$ hence

$$I = E_f(w(x)),$$

$x \sim Uniform(a,b)$. If we generate $X_1, X_2, ...X_n \sim Uniform(a,b)$ where N is large, the law of large numbers gives us:

$$I = \frac{1}{N}\sum_{i=1}^{n} w(X_i) \xrightarrow{p} E(w(x)) = I.$$

Applying this method to our problem we have:

$$\int_a^b h(x)dx = \int_a^b w(x)f(x)dx = \int_T P(D|T,\theta)P(T|\theta)dT \cong \frac{i}{N}P(D|T_{(i)},\theta),$$

where $T_{(1)}, T_{(2)}, ..., T_{(N)} \sim P(T|\theta)$.

Unfortunately this rather straightforward method has one major drawback it is quite inefficient for large data sets. However both IS and MCMC methods improve the efficiency of basic Monte Carlo integration.

2.1.2 Importance Sampling

When trying to evaluate

$$I = \int_a^b h(x)dx,$$

the basic Monte Carlo method involves sampling from f, which in our problem above is $P(D|T,\theta)$. By attempting to concentrate on computing trees for which $P(D|T,\theta)$ is large, the computational effort involved can be reduced. I can be rewritten as follows:

$$I = \int w(x)f(x)dx = \int \frac{w(x)f(x)}{g(x)}g(x)dx = E_g(Y),$$

where $Y = \frac{w(x)f(x)}{g(x)}$. Simulating $X_1, X_2, ...X_n \sim g$ for N sufficiently large, I can be estimated approximately by

$$I = \frac{1}{N}\sum_i Y_i = \frac{1}{N}\sum_i \frac{w(X_i)f(X_i)}{g(X_i)},$$

and by the law of large numbers $\hat{I} \xrightarrow{p} I$. Although we are still left with the same problem of sampling from an unknown distribution, we now have a proposal distribution g, which in our case we shall denote as $Q(T)$. If $Q(T)$ is chosen wisely, the technique will simulate trees that contribute significantly to the likelihood and therefore the computation time is reduced. The optimal choice Q^*_θ for Q() is the post data distribution of the tree T given the sample data and θ and is

$$Q^*_\theta(T) = P(T|D,\theta) = \frac{P(T|\theta)P(D|T,\theta)}{P(D|\theta)} = \frac{P(T,D|\theta)}{P(D|\theta)}.$$

The fact that this method focuses on attempting to estimate $L(\theta)$ by concentrating on "informative" or "important" trees yields the name Importance Sampling and $Q(T)$ is referred to as the Importance Sampling distribution.

2.1.3 Markov Chain Monte Carlo (MCMC)

Returning once again to our initial problem, we wish to generate trees from some distribution which is difficult to simulate from. Importance sampling offers one way of getting around this problem and MCMC offers an alternative approach. The idea in the MCMC approach is to construct a Markov Chain $X_0, X_1, ...$ whose stationary distribution is f. We then have that

$$\frac{1}{N}\sum_{i=1}^{N} H(X_i) \rightarrow E_f(h(X)) = I,$$

under certain conditions. There are several MCMC techniques that are in use e.g. The Metropolis-Hastings Algorithm, Gibbs Sampling and Accept-Reject Sampling. However the Metropolis-Hastings Algorithm is the most widely used method and for this reason is the one we shall explain in detail.

The Metropolis-Hastings Method

The method works as follows: let Q denote an arbitrary distribution that we are able to sample from (in our case Q will move randomly from tree to tree). The Metropolis-Hastings algorithm ensures that in the long run, the sequence of observations sampled from Q, *i.e.* $X_0, X_1, ...$ will represent a Markov Chain with f as its stationary distribution.

The following steps lead to this:

1. Choose X_0 arbitrarily.

2. Given X_i, which has been generated from $X_0, X_1, ..X_{i-1}$, generate a proposal value $Y \sim Q(y|X_i)$

3. Evaluate A where

$$A = min\left(\frac{f(y)Q(x|y)}{f(x)Q(y|x)}, 1\right)$$

4. Set $X_{i+1} =$

$$\begin{cases} Y & \text{with probability A,} \\ X_i & \text{with probability 1-A.} \end{cases}$$

Now, a sample of observations that approximate a sample from the distribution f has been constructed. Additionally, for k sufficiently large, $X_0, X_{0+k}, X_{0+2k}, \ldots$. may be considered independent samples from f. A good choice of Q will increase the efficiency of the algorithm whiles a bad choice may render the method ineffective. It is often difficult to predict what proposed distribution will be the best one *a priori* but there are some pointers to look for during simulations. If Q is found to propose high values for the distribution f then it is considered a good distribution, otherwise almost every proposed value for X will be rejected and the algorithm will remain at the same value for a long time. Also, some proposed distributions will result in small changes being made over a long period of time. This is not a good property and in fact a distribution that does the opposite *i.e.* an algorithm that moves freely between all possible values of X is desirable. To sum up, a proposal distribution Q which moves through different values of X quickly and which also has a high acceptance rate (*i.e.* high values for A) is an ideal candidate. This can be found by simply trying out different distributions until a suitable one appears.

2.2 MapArg

MapArg, the fine mapping method developed by Larribe (2002), is of primary interest to us throughout the rest of this body of work. In fact, the methods we develop to correct for incomplete penetrance and phenocopy (described in Chapter 3) are then incorporated into the MapArg framework to see if adjusting for these parameters can improve the performance of this fine mapping method. It is important, therefore, to describe the theory behind MapArg in some detail.

2.2.1 The Model

Consider a sample of sequences where the cases have the status affected and the controls non-affected. The status of the phenotype is caused by a single mutation that occurs in the history of the population. This model is known as the infinite sites model. It is assumed that the infinite sites model describes a sequence of DNA with an infinite amount of loci. The mutation rate is so small at each position that it is fairly realistic to suppose that only one mutation occurs at one locus throughout the population history. It is also assumed that the population is young and isolated.

There are L markers in a sequence, $L-1$ of which are known. Locus m represents the position of the marker m in the sequence. The TIM is also a marker in this sequence but its exact position is unknown. The methods proposed are intended to estimate the location of the TIM. Letting r_T denote the distance between the start of the sequence and the position of the TIM, an estimator for r_T is obtained by maximum likelihood methods.

Let r be the total length of a sequence and let x_m denote the position of marker m. Without loss of generality, say that the first marker starts at the origin, giving:

$$\begin{cases} x_1 = 0, \\ x_m = \sum_{p=1}^{m-1} r_p \quad 2 \leq m \leq L. \end{cases}$$

When a coalescent, recombination or mutation event occurs at time t, the ancestral material of the sequences at the L loci are affected. Denote the set of ancestral sequences at time t_τ, by H_τ where τ ranges from 0 to τ^*, 0 representing the sequence at the present time and τ^* being the last coalescent that results in the MRCA. The following notation will be used to describe the events that can occur when going back in time from 0 to τ^*:

$$\begin{cases} C_i & \text{Coalescence of two identical sequences } i \\ C_{ij}^k & \text{Coalescence of sequences } i \text{ and } j \text{ into sequence } k \\ M_i^j(m) & \text{Mutation of sequence } i \text{ into sequence } j \text{ at marker.} m \\ R_i^{jk}(p) & \text{Recombination of sequence } i \text{ into sequences } j \text{ and } k \text{ in the interval } p \end{cases}$$

The probability distribution for H_τ, represented by $Q(H_\tau)$ and $Q(H_\tau)$ is a function of $Q(H_{\tau+1})$. If a recombination event occurs at time $\tau + 1$, $H_{\tau+1}$ is written as $H_{\tau+1} = H_\tau + R_i^{jk}(p)$. Similarly, the occurrence of a coalescent event on step in the past, is denoted by $H_{\tau+1} = H_\tau + C_i$, when two identical sequences coalesce and $H_{\tau+1} = H_\tau + C_{ij}^k$ when two distinct sequences coalesce. Finally, a mutation that occurs one step in the past is denoted $H_{\tau+1} = H_\tau + M_i^j(m)$.

If a recombination event occurs the distribution for the point of recombination is:

$$f_z(z) = \frac{1}{r}, \quad \text{if } 0 < z < r. \tag{2.1}$$

A recombination event can happen anywhere along the sequence but the only place that such an event effects the ARG is when a recombination occurs on the part of the sequence that is ancestral. Consider two sequences s^1 and s^2, each with 5 markers. The five markers on s^1 are ancestral whereas markers 1 and 2 of s^2 ancestral and markers 3, 4 and 5 are non ancestral. Say that the recombination event occurs between locus 2 and locus 3 which results in two parental sequences: a left sequence containing marker 1 and 2 and a right sequence containing marker 3, 4 and 5. For s^1, the history of the sample will be modified as both the left and right parental sequences contain ancestral material. However the history of the sample for s^2 remains unchanged since the left parental sequences is similar to the original sequence and the right parental sequence contains only non ancestral sequence which does not change the sample history. Define c_i as the proportion of sequence i for which a recombination event can effect the ancestral material and let b represent the total length of all sequences where a recombination event can effect the ancestral material for H_τ, where $0 \leq b \leq nr$. Also, by the same reasoning, a mutation event will only effect the sample history if the event occurs at an

ancestral marker. We define a as the number of markers out of all H_τ for which the sample history can be effected by a mutation event and $n \leq a \leq nLs$. The coalescence of two different sequences i and j is possible if the two sequences are similar apart from the non ancestral segments. In figure 2.1 for example, s^1 and s^2 are two different sequences but they only differ at marker 3, where s^1 possesses the ancestral wild type allele and s^2 non-ancestral material. However, if a sequence i has a mutant allele at one marker and sequence j does not have the mutant allele for the same marker but carries the ancestral wild type marker, then they are unable to coalesce. Looking at the sequences in figure 2.1, we can see that s^1 and s^3 are unable to coalesce since the third marker for s^1 carries the wild type allele and the third marker for s^3 carries the mutant allele.

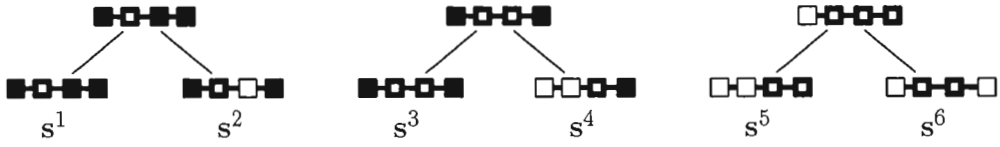

Figure 2.1 Examples of possible coalescent events between different sequences. Shaded-in boxes represent ancestral wild type alleles, un-shaded boxes represent non-ancestral material and half-shaded boxes an ancestral mutant allele (from Larribe, 2003).

2.2.2 The Probability of a Mutation, Coalescent or Recombination Event

Now that the parameters of the model have been introduced, the probabilities of the three possible event in the ARG, namely a mutation, the coalescence of two sequences or a recombination event, are developed in detail. Once these probabilities are determined they are incorporated into the probability distribution $Q(H_t)$.

We shall work with a sample of size n, where there are n_i sequences of type i for $i = 1, ..., d$. The rate of mutation per sequence per generation for this sample is μ and

the total rate of recombination is ρ. The time taken for the first coalescence between two sequences to occur, T_C, follows an exponential distribution with mean $\frac{n(n-1)}{2}$ (See section 1.3.4). Since μ is the rate of mutation per sequence per generation, the rate of mutation per generation is $4N\mu = \theta$. It can be shown in a similar fashion to that of the coalescent case, that the time taken for a mutation event to arise follows an exponential law with mean $\frac{n\theta}{2}$, denoted by T_M. Also the time taken for a recombination, T_R to happen follows an exponential distribution with mean $\frac{n\rho}{2}$. Summarizing:

$$T_C \sim \text{Exp}(\lambda_C) = \text{Exp}\left(\frac{n(n-1)}{2}\right),$$
$$T_M \sim \text{Exp}(\lambda_M) = \text{Exp}\left(\frac{n\theta}{2}\right),$$
$$T_R \sim \text{Exp}(\lambda_R) = \text{Exp}\left(\frac{n\rho}{2}\right).$$

What we are really interested in calculating is the following: given that an event has occurred in the past, what is the probability that it is a coalescent event, a mutation even, or a recombination event *i.e.*

$$P(Co|\text{Co or Mu or Re}),$$

$$P(Mu|\text{Co or Mu or Re}),$$

$$P(Re|\text{Co or Mu or Re}).$$

If X_1, X_2, X_3 are independent exponential random variables with respective rates λ_1, λ_2 and λ_2, then the following is true:

$$P(X_i = \min_j X_j) = \frac{\lambda_i}{\sum_{j=1}^{3} \lambda j}.$$

In our context this gives:

$$P(Co|\text{Co or Mu or Re}) = \frac{n(n-1)}{n(n-1)+n\theta+n\rho},$$
$$P(Mu|\text{Co or Mu or Re}) = \frac{n\theta}{(n-1)+n\theta+n\rho},$$
$$P(Re|\text{Co or Mu or Re}) = \frac{n\rho}{n(n-1)+n\theta+n\rho}.$$

If a coalescence occurs then there are $n-1$ sequences one step in the past. Two identical sequences i coalesce with probability $(n_i - 1)/(n-1)$. If two different sequences i and j coalesce to sequence k in the past, we have to account for the fact that k may be identical to sequence i or j or even neither sequence. Then, the probability that i and j coalesce to sequence k is $(n_k + 1 - \delta_{ik} - \delta_{jk})/(n-1)$, where $\delta_{jk} = 1$ if $i = k$ and 0, if $i \neq k$.

Suppose that the first event is a mutation, then there exists a sequence i that comes from sequence j, that may already exist or not, with probability $(n_j + 1)/n$. Note that the number of sequences one step in the past do not change if the event occurring is a mutation. Given that the mutation rate θ is the rate for the whole sequence, the probability of a mutation at a given marker it thus $(n_j + 1)/nL$. Furthermore, a mutation on the nonancestral material occurs with probability $(nL - a)/nL$.

A recombination event can occur anywhere on the sequence and as seen in section 2.1.1, if the event occurs in a given interval on sequence i, the left parental sequence j shares the same genetic material as i to the left of the point of recombination whereas the right parental sequence k shares the same genetic material as i to the right of the point of recombination. One step back in time, we have $n_j + 1$ sequences of type j and $n_k + 1$ sequences of type k and the total number of possible ordered pairs or sequences is $n(n+1)$ (a recombination event results in one more sequence in the sample history one step in the past). Furthermore, the probability that a recombination event occurs in a certain interval, is proportional to the length of this interval. So, the probability that there is a recombination in interval p is r_p/r. Thus the probability of a recombination event is $[r_p/r].[(n_i + 1)(n_k + 1)]/n(n+1)$. The combination of these facts lead to a recursion initially introduced by Griffiths and Marjoram (1996). MapArg uses an analogous version of this recursion in order to represent the probability distribution $Q(H_\tau)$.

2.2.3 Recurrence Probability for the distribution of Q()

Bearing in mind we are looking at the state of H_τ which depends on $H_{\tau+1}$, the probability distribution can be written as follows:

$$
\begin{aligned}
Q(H_\tau) &= \frac{n(n-1)}{D}\Bigg[\sum_1 \frac{n_i-1}{n-1}Q(H_\tau + C_i) \\
&+ 2\sum_2 \frac{n_k+1-\delta_{ik}-\delta_{jk}}{n-1}Q(H_\tau + C_{ij}^k)\Bigg] \\
&+ \frac{n\theta}{D}\Bigg[\sum_3 \frac{n_j+1}{n}Q(H_\tau + M_i^j) \\
&+ \frac{nL-a}{nL}Q(H_\tau)\Bigg] \\
&+ \frac{n\rho}{D}\Bigg[\sum_{i=1}^{d}\Bigg(\sum_{p=\gamma_i}^{\kappa_i}\rho_p\frac{(n_j+1)(n_k+1)}{n(n+1)}Q(H_\tau + R_i^{jk}(p))\Bigg) \\
&+ \frac{nr-b}{nr}Q(H_\tau)\Bigg],
\end{aligned}
$$

where D = $[n(n-1)+n\theta+n\rho]$ each line refers to the following events:

Line 1 Coalescence of two sequences of type i,

Line 2 Coalescence of sequence i and j to sequence k, where $i \neq j$,

Line 3 Mutation of sequence i to sequence j, where j may already exist,

Line 4 Mutation in non ancestral material,

Line 5 Recombination of sequence i, in interval p, that produces sequences j and k, where j and k may already exist,

Line 6 Recombination of non ancestral material,

and the numbers under the summation signs mean:

(1) Summation over sequences $i = 1, ..., d: \quad i; n_1 > 1$,

(2) Summation over unordered pairs i, j that possess the same set of mutations in the ancestral material,

(3) Summation over all singleton mutations.

The above equation can look quite daunting at first glance, but the essential thing to note is that it is a recursive equation that accounts for all possible events that can occur one step back in the generation from which you start at. Thus, starting at time $\tau = 0$, *i.e.* the haplotype data in the present generation, the ARG can be constructed one step at a time, from $\tau = 0$ until the point where all sequences join to give the MRCA of the sample at time τ^*.

2.2.4 Importance Sampling Within The MapArg Framework

Genealogies are produced at each interval along the sample sequence, and they are generated using the above recursive equation. As can be seen however, there are several parameters to estimate at once and so simulation techniques are employed as it is not possible to find an exact solution for the equation. MapArg uses Importance Sampling techniques, one of two popular Monte Carlo simulation methods in use by fine mapping researchers. Section 2.1.2. gives a general explanation of the ideas behind Importance Sampling (IS), and here we shall describe how it is employed within the MapArg framework.

Define a Markov Chain with transition probabilities from H_τ to $H_{\tau+1}$, denoted by $P(H_{\tau+1}|H_\tau)$. Now $\tau\epsilon(0, ..., \tau^*)$ and the chain reaches it's absorbing state then a common ancestor is found for all sequences, at time τ^*. As the genetic code of the ancestor is supposed to be known, then $Q(H\tau^*)$ is 1 for a single sequence and 0 for all others. From Section 2.2.3, we see that the recurrence equation could be written as:

$$Q(H_\tau) = \sum_{H_{\tau+1}} Q(H_\tau|H_{\tau+1})Q(H_\tau + 1),$$

but we want to build graphs from the present to the MRCA so we insert a proposal distribution in the recurrence equation:

$$Q(H_\tau) = \sum_{H_{\tau+1}} \frac{Q(H_\tau|H_{\tau+1})}{P(H_{\tau+1}|H_\tau)} P(H_{\tau+1}|H_\tau) Q(H_{\tau+1}).$$

Defining

$$f(H_\tau, H_{\tau+1}) = \frac{Q(H_\tau|H_{\tau+1})}{P(H_{\tau+1}|H_\tau)},$$

the recurrence equation can be written in the following form:

$$Q(H_\tau) = \sum_{H_{\tau+1}} f(H_\tau, H_{\tau+1}) P(H_{\tau+1}|H_\tau) Q(H_{\tau+1}).$$

In particular,

$$\begin{aligned} Q(H_0) &= \sum_{H_1} f(H_0, H_1) P(H_1|H_0) Q(H_1) \\ &= \sum_{H_1} f(H_0, H_1) P(H_1|H_0) \left[\sum_{H_2} f(H_1, H_2) P(H_2|H_1) Q(H_2) \right] \\ &= \cdots \\ &= \sum_{H_1} \sum_{H_2} \cdots \sum_{H_{\tau^*}} f(H_0, H_1) f(H_1, H_2) \cdots f(H_{\tau^*-1}, H_{\tau^*}) * \\ &\quad P(H_0|H_{-1}) P(H_1|H_0) P(H_2|H_1) \cdots Q(H_{\tau^*}) \end{aligned}$$

and therefore

$$Q(H_0) = E_P \left[\prod_{\tau=0}^{\tau^*-1} f(H_\tau, H_{\tau+1}) \right]. \tag{2.2}$$

This is an importance sampling representation where $P(H_{\tau+1}|H_\tau)$, is known as the proposal distribution for $Q(H_\tau)$. Let $\Theta = (\theta, r_T)$ be the set of unknown parameters of the process. Then given $\Theta_0 = (\theta_0, r_{T_0})$ an estimate of $Q_\Theta(H_\tau)$ can then be found for different values of Θ, by a second importance sampling procedure (details not given here). This method evaluates the likelihood of r_T along a sequence of samples with L-1 known markers. In order to run simulations a driving value must be provided for each interval p, which in the case of MapArg is taken to be the middle of the interval p where $1 \leq p \leq L-1$. Then the IS technique described above constructs graphs with the driving value and a likelihood for each region (x_p, x_{p+1}) is evaluated. The maximum likelihood estimate is then taken as the estimate for r_T, the position of the TIM.

2.2.5 Composite Likelihood

When calculating the likelihoods in the above equation, the simulations of the genealogies may contain a lot of variability and so depending on the sample size and the number of markers involved the computation time can take days even weeks. Larribe (submitted 2007) then introduced composite likelihood in order to control for variation. The idea behind composite likelihood, as the very name suggests, is to divide the set of markers into small sections or windows. Figure 2.2 shows some examples of windows using d contiguous observed markers.

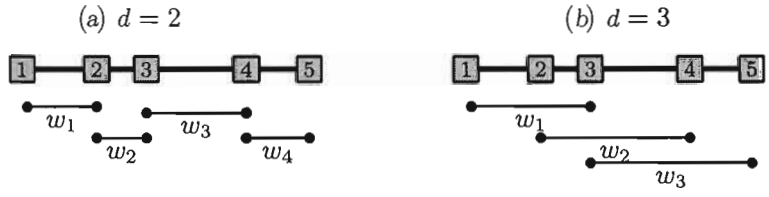

Figure 2.2 An illustration of window lengths for a sequence of 5 markers, where d equals the window length (from Larribe, 2007).

Consider a sequence with L markers, $L-1$ of which have a known position, then $G = L-d$ is the total number of windows for this sequence. The interval m, located between the $m-1$th and mth marker, will be included in the window g, for $g = (1, ..., G)$, if and only if g is between $\underline{g}(m)$ and $\overline{g}(m)$ where

$$\underline{g}(m) = max(1, m+1-d)$$
$$\overline{g}(m) = min(m-1, L-d).$$

Let $L_{m,g}(r_T)$ represent the marginal likelihood function for the position of the TIM in the interval m. The likelihood is evaluated using the information on the markers in window g only. A composite likelihood function giving equal weight to each window can be defined as

$$CL_d(r_T) = \prod_{m=2}^{L-1} \left(\prod_{g=\underline{g}(m)}^{\overline{g}(m)} L_{m,g}(r_T) \right)^{w_m},$$

where

$$w_m = \frac{1}{\underline{g}(m) + \overline{g}(m) + 1}.$$

This function is not well defined however, since $L_{m,g}(r_T)$ uses different information from on window, g, to the next. A likelihood that is conditional on the d markers in window g, denoted H_0^g, is thus proposed. This conditional likelihood is defined as

$$L_{m,g}(r_T|H_0^g) = \frac{Q(H_0^{r_T}, H_0^g)}{Q(H_0^g)},$$

where $H_0^{r_T}$ represents the sample configuration at the TIM including its position r_T, in interval m. The corresponding composite-conditional-likelihood is

$$CCL_d r_T = \prod_{m=2}^{L-1} \left(\prod_{g=\underline{g}(m)}^{\overline{g}(m)} L_{m,g}(r_T|H_0^g) \right)^{w_m}.$$

When windows of size d are used, the likelihood function is estimated $d-1$ times in all intervals apart from the first and last $d-2$ windows. Looking at the example in Figure 2.2 (b), $d = 3$ and $L = 6$. In the first and last interval the likelihood function is estimated once only, whereas for intervals 2 and 3, the likelihood is estimated twice. If the number of markers is large compared to the window size then almost all of the intervals estimate the likelihood function $d-1$ times. Results have been produced for different window lengths and it appears thus far, that the method is efficient at finding the location of the TIM and hence MapArg uses composite likelihood in estimating r_T.

2.3 Other Fine Mapping Methods

Before developing a method that corrects for incomplete penetrance and phenocopy within the MapArg framework, let us discuss some of the other fine mapping methods that are also based on the coalescent model. It is of particular interest to see how these other methods account for the penetrance parameters.

2.3.1 Decay of Haplotype Sharing

Decay of Haplotype Sharing (DHS) is a model that was developed by McPeek and Strahs (1999). It is based on the idea that there is a certain segment surrounding the TIM locus that is common to all cases originating from the MCRA or founder, that reduces in size or "decays" over generations. By this hypothesis, any position on the chromosome that is well explained by decay of shared haplotype is likely to be a candidate for the location of the TIM. As the name suggests DHS works with haploid data *i.e.*, chromosomes are considered as independent sequences and not in pairs .

The authors start off by looking at an individual haplotype. They model the complete ancestral history of this haplotype and then generalize to the complete sample of sequences. This generalization is done differently depending on the situation. There are two different cases that are modeled namely:

1. The case where independence across haplotypes is assumed,
2. The case where there is a dependency or correlation between different haplotypes.

The genetic distance from the location of the TIM, r_T to the end point of the segment where the haplotype still shares ancestral data with the founder is denoted x. A function $R(x)$ is then defined as A (Ancestral) for any x, a distance from the TIM locus, that shares ancestral material with the founder and is N (Non Ancestral) otherwise. The function $R(x)$ it then considered a continuous time Markov chain that is indexed by position with

$$Pr[R(x+\delta) = A|R(x) = A], \quad Pr[R(x+\delta) = N|R(x) = N],$$

for x and $t > 0$. Then, a parameter for estimating the r_T is introduced so the likelihood estimate now estimates both x and r_T simultaneously. In order to reduce computation time, the likelihood is maximized on a set of candidate values for x. The candidate values are obtained by carrying out a branch-and-bound procedure (Baum, 1972). Other parameters are introduced into the model to deal with chance sharing of alleles and

to account for multiple origins of the TIM. The latter phenomenon is of interest to us and shall be explained in more detail shortly. When DHS is modeled assuming independence of sequences, the likelihood of the whole sample is obtained by multiplying each individual haplotype likelihood. If on the other hand this assumption is not valid a more complex method is employed. Using a sample of unrelated individuals, McPeek and Strahs estimated r_T with a quasi-likelihood estimate. The usefulness of this being, it resembles the likelihood estimate above but takes into account correlation between haplotypes. The authors have shown that when using a quasi-likelihood estimating equation under the coalescent model that the same estimate as the independent case is obtained but the standard error is larger.

Multiple origins of the TIM can be viewed as a form of phenocopy. Recall the definition of phenocopy:

$$\text{Phenocopy} = f_0 = Pr(\text{Affected}|\overline{TIM})$$

MapArg searches for a single position on the chromosome that causes a disease, *i.e.* the TIM locus. Consider the situation where there are two different loci that have different TIMs , T_1 and T_2 and T_1 accounts for a large proportion of cases among the population. The fine mapping methods discussed will pick up the LD surrounding T_1 and produce an estimate for the location. However T_2 has an allele that accounts for a small proportion of cases also. Within the MapArg framework the cases that result from the TIM at T_2 will be viewed as a phenocopies as they are sequences that have a mutation but don't share the same alleles as the set of cases stemming from the locus T_1. McPeek and Strahs introduce a parameter p that represents the proportion of cases in the population that are nor descended from the founder haplotype, and $1-p$ is the proportion that does descend from the founder. Then the likelihood is then calculated conditional on the sample haplotype being a case. A simplification of their model is

$$L(TIM) = (1-p).L(r_T, x|D) + p.P_{null}(D).$$

where $P_{null}(D)$ is the probability of the sample data estimated from the control population. When $p = 0$ then there is no other origin of the TIM considered. For the

simulations that were run in McPeek and Strahs (1999) the model still seems to perform well then p is non zero.

2.3.2 Fine Mapping Via The Shattered Coalescent

Another variation on the coalescent model for estimating the position of the TIM is the shattered coalescent model. Morris *et al.* (2002) developed the idea of modeling genealogies that may split into smaller sections that represent multiple founding mutations or sporadic cases of the disease. Their fine mapping method is implemented with Markov Chain Monte Carlo techniques (See Section 2.1.3). A Bayesian approach is taken and the authors approximate a posterior probability distribution for r_T, conditional on the sample data. As the shattered coalescent accounts for phenocopy it shall be described in detail. An overview of the work of Morris *et al.* (2002) shall be given first. Once the shattered coalescent is modeled and incorporated into the method the genealogies are constructed maximizing several parameters needed to generate trees at a given candidate locus simultaneously. As in all fine mapping methods of multi locus data the computation is intensive and so an algorithm similar to the Metropolis-Hastings one (see section 2.1.3) is developed. The primary parameter of interest in the MCMC simulation is r_T but several other parameters are also estimated during the process of generating genealogies including allele frequencies and LD parameters.

The Shattered Coalescent can be viewed as a generalization of the coalescent process where branches of the genealogical tree can be removed. The likelihood for the controls is considered to depend on the control haplotypes only. So the shattered coalescent models cases only and not controls. Let

$$z_b = \begin{cases} 1, & \text{if node } b \text{ has a parental node in the shattered tree ;} \\ 0, & \text{if node } b \text{ has no parental node in the shattered tree,} \end{cases}$$

where the node can be internal or a leaf, where leaf nodes represent the sample haplotypes at the present generation and internal nodes the sample haplotypes at every other point in time in the history of the sample. If there is a leaf node with $z_b = 0$

this represents a phenocopy. The reason is, it represents a haplotype that has no common ancestors with the rest of the sample, which means that it is a sequence that is effected by some disease without having the TIM. Now consider node b that has marker haplotype C_b. Two scenarios are considered:

1. Node b has no parent node in the underlying genealogy. Then $z_b = 0$ and node b corresponds to either a founder for a disease mutation at x when the node is internal, or a sporadic case or phenocopy. Founding mutations and phenocopies are assumed to occur on random chromosomes from the population and thus are modeled in the same way as control haplotypes. The simplest model for controls being one assuming no LD in which case the likelihood is given by the product of population proportions for each allele.

2. Node b has a parent node in the underlying genealogy and z_b =1. The distribution of C_b now depends on the haplotype of the parental node, the occurrence of recombination and also mutations along the branch connecting C_b and the parent haplotype.

Recombination and mutation events, ρ and μ are assumed to occur independently across the branches of the tree. The overall likelihood is of the form:

$$L(TIM) = \prod_b [L(C_b|P_b, x, N, h, \rho, \mu)z_b + L(Cb|h)(1 - z_b)],$$

where P_b represents the haplotype at the parental node, N the population size and h the haplotype data. Any unknown parameter is also estimated within the MCMC simulation framework. It can be seen from the above equation that haplotypes that are thought to be phenocopies are treated similarly to the control haplotypes of the sample.

2.3.3 Tree LD

The underlying model for the methods of Zöllner and Pritchard (2005) is also the coalescent, and to take recombination into account an adaptation of the ARG is used. The

authors' method is implemented by executing a program called TreeLD and for this reason we shall refer to their fine mapping method as TreeLD from now on. Zöllner and Pritchard work with haplotypes also, but a distinct difference between TreeLD and the methods of Morris *et al.* and McPeek and Strahs is that the genealogies constructed within the coalescent framework are done so by using both case and control information. This means that more information is being used which can improve the accuracy of estimating r_T, but it also means that computation will be a lot more involved than when just using the cases to construct genealogies. In order to try and reduce the computation time they construct the genealogies of the n sample haplotypes independently of the phenotype data and mutation status and then obtain an likelihood of the phenotypes conditional on the genealogies and mutation rate: *i.e.* at a given candidate position for the TIM, x, the sample history is constructed for the entire sample of cases and controls, after which the mutation rate is superimposed on the genealogy and the distribution of the phenotype data is obtained. In estimating the location of the TIM a Bayesian approach is adopted using MCMC techniques. As in the DHS method, TreeLD constructs the genealogies on a grid of candidate loci across the sample sequence. The Metropolis-Hastings algorithm is employed to construct genealogies at each candidate locus, x, within the simulation. As mentioned before, the phenotype likelihood is calculated after the trees are simulated so the distribution being modeled within the MCMC simulations is $P(T_x|x, G)$, where T_x represents the genealogy at x and G is the genotype data or haplotypes. Once the distribution $P(T_x)$ is found, a peeling algorithm (Felsenstein, 1981) gives the probability distribution for the phenotype data at x. Then the posterior distribution given as:

$$P(x|\Phi, G) \simeq P(\Phi|x, G) \cdot P(x). \tag{2.3}$$

The prior distribution $P(x)$ is taken as uniform across the entire region of the sample but can be modified accordingly and $P(\Phi|x, G)$, the probability distribution of the phenotypes, will be explained further below. Equation (2.3) is the general form of a posterior distribution from standard theory on Bayesian Inference. Once $P(x|\Phi, G)$ is obtained, the mode is taken to be the estimate of the location of the TIM locus.

Since TreeLD uses haploid data, the penetrance parameters are defined as they are for MapArg. Recall that

$$\begin{aligned} \text{Penetrance} &= f_1 = Pr(\text{Affected}|TIM) = Pr(S_i = 1|M_i = 1), \\ \text{Phenocopy} &= f_0 = Pr(\text{Affected}|\overline{TIM}) = Pr(S_i = 1|M_i = 0). \end{aligned}$$

The model for the distribution of phenotypes conditional on the genealogies ($P(x|\Phi, G)$) assumes that the disease mutation occurs as a Poisson process with rate $v/2$ and that multiple mutations on the same chromosome have no further effect: every chromosome that carries at least one mutation has the same distribution. Let K_j, $j\epsilon(1, 2, ..., 2n-1)$ be the ordered set of nodes on the genealogy, T_x, so that $K_1, ..., K_n$ are the external nodes or leafs, K_{n-1} is the node of the first coalescent event and K_{2n-1} is the MRCA. Furthermore, let $B = (b_1, ..., b_{2n-2})$ be the vector of branch lengths where b_j is the branch length between node K_j and its parental node. Denote m_i as an indicator variable for the mutation status at node K_i where

$$m_i = \begin{cases} 1, & \text{if node } K_i \text{ carries at least one mutation;} \\ 0, & \text{otherwise.} \end{cases}$$

Now if there is a mutation at node K_i, then all phenotypes that are a descendants of this node will be cases. Likewise, phenotype descending form ancestral nodes without a mutation will be controls. Therefore,

$$P(\Phi_i|m_i = 1) = (f_0)^{n_{controls}}.(f_1)^{n_{cases}},$$

where Φ_i represents the phenotypes of all leaf nodes that descend form node K_i.

To calculate $P(\Phi_i|m_i = 1)$, it must be taken into account that given $m_i = 0$ at node K_i, the ancestral nodes may or may not have the status $m_i = 0$, since a mutation could have occurred on the branch. This is done by using the information that the mutation rate $v/2$ occurs according to a Poisson process. It is therefore possible to calculate the probability of every node by starting at the most recent nodes and working iteratively backwards in time. It is clear though, that to calculate $P(x|\Phi, G)$ with the peeling algorithm above, there has to be values given for f_1 and f_0. Zöllner and Pritchard take

these penetrance parameters as being any value of the bounded set $\delta = [0,1] * [0,1]$. The likelihood of the phenotype is calculated for various values of δ and the values that maximize this likelihood are then used as the penetrance parameters f_1 and f_0 for the data. Note that TreeLD accounts for incomplete penetrance as well as phenocopy which was not done in the previous literature.

CHAPTER III

ACCOUNTING FOR INCOMPLETE PENETRANCE AND PHENOCOPY

3.1 Effects of Penetrance and Phenocopy

3.1.1 Introduction

Penetrance and Phenocopy have been introduced in Chapter 1 (see section 1.4) and a small example of their effect on the ability to distinguish people possessing a TIM from those who don't was shown. It is of interest to determine exactly how various levels of penetrance and phenocopy are expected to influence the efficacy of MapArg in producing an estimate of r_T. Let M_i denote status of sequence i, $(i = 1, ..., n)$, where $M_i = 1$ when the sequence contains the TIM and 0 otherwise. Let S_i denote the affected status of phenotype for individual i, where $S_i = 1$ for cases and 0 for controls. The parameters are defined as in Section 1.4 but since the notation has been shortened for convenience we shall present them again:

$$\text{Penetrance} = f_1 = Pr(\text{Affected}|TIM) = Pr(S_i = 1|M_i = 1), \tag{3.1}$$

$$\text{Phenocopy} = f_0 = Pr(\text{Affected}|\overline{TIM}) = Pr(S_i = 1|M_i = 0). \tag{3.2}$$

3.1.2 The Structure of the Data

Initially MapArg needs a data set in a certain format along with certain information on this data to execute the computer program. The structure of the data along with

parameter information is as follows:

- A sample of SNPs (haplotypes), of size n, from the population of interest;
- The frequency of each haplotype in the sample;
- A classification of the individuals in the sample, in other words, each sequence is classified as being either a case or a control;
- The distance between markers.

Here is a example of the type of data with which we can implement MapArg. There are 16 sequences, genotyped at 4 loci *i.e.* 4 SNPs of which 7 are cases and 9 are controls.

Sequence	Case	Control
0001	1	5
0011	2	3
1101	4	1

Table 3.1 An example of the data MapArg works with. The first column represents the sequence in haplotype form, and the other two columns contain the frequency of cases and controls for each haplotype.

With this information we are able to distinguish cases from controls. MapArg has assumed complete penetrance and no phenocopies, *i.e.* $f_1 = 1$ and $f_0 = 0$ until now. If the TIM in question is such that $f_1 < 1$ and $f_0 > 0$, we would like to be able to take this into account when estimating the position of the TIM along the section of chromosome being studied. Before the development of a model to correct for the penetrance parameters we shall look at how incomplete penetrance and phenocopies effect the performance of MapArg.

3.1.3 Simulating data with varying levels of Penetrance and Phenocopy

A program called SimPenPhen, allowing us to simulate data with various levels of f_1 and f_0 was written in C++. One should note that there is no theoretical way to study these effects on the mapping method we are working on, MapArg. A solution is to simulate some examples where the result is known (*i.e.* the real position of the TIM), under various conditions, and to observe the effect. By taking several random examples, one hopes to better understand and illustrate what one is unable to derive directly from mathematics.

Suppose that we have good reason to believe that $f_1 = 0.85$ and $f_0 = 0.17$. SimPenPhen takes a sample data set of the form given in table 3.1, and outputs a data set that represents data with a penetrance of 85% and 17% phenocopy. Then instead of working with a sample of sequences that are considered cases or controls, we will work with a sample of sequences that are considered mutant, if carrying the TIM or non-mutant otherwise. Continuing the example in Section 1.4 (see page 16), we assume that the input data for SimPenPhen is as follows:

Sequence	Case	Control
0001	5	9
0011	2	15
1100	6	12
1101	7	4

Table 3.2 A data set with 5 sequences genotyped at 4 loci, *i.e.* 4 SNPs per sequence. There are 20 cases and 40 controls in total.

Given $f_1 = 0.85$ and $f_0 = 0.17$, one realization of executing SimPenPhen is as in Table 3.3. Note that the amount of sequences for each haplotype remains unchanged e.g. haplotype "0001" has 14 sequences in the data set. However MapArg will use data that has 24 mutants and 36 non-mutants as opposed to 20 cases and 40 controls, which

Sequence	Mutant	Non-Mutant
0001	6	8
0011	4	13
1100	7	11
1101	7	4

Table 3.3 An example of the output data given from SimPenPhen when $f_1 = 0.85$ and $f_1 = 0.17$. There are 24 mutant and 36 non-mutant sequences in total.

would have been the case if the data were not adjusted to account for the penetrance parameters.

Here are the main steps that the program executes in order to simulate data sets having the same form as table 3.3. In section 3.4 below, we will see how the original samples of sequences (*i.e.* data of the form in table 3.2) we use as input data for SimPenPhen are simulated. Let's suppose for now that we have such a sample.

Main steps involved in the C++ Program, SimPenPhen

1. The input data is a "population" of size 10 000; a given marker is chosen to be the TIM, and this information is put aside.

2. For each of these 10 000 sequences, if a haplotype is a mutant (as determined by the chosen marker in step 1), then it has a probability of f_1 of being a case, and $1 - f_1$ of being a control. If a haplotype is a non-mutant, then it has a probability of f_0 of being a control, and $1 - f_0$ of being a case. A random number is generated and according to this random number, we "transfer" mutant/non-mutant information to case/control status.

3. From this population, a sample is chosen conditional on the the disease status; for example, 100 cases and 100 controls.

Since we already know the real position of the TIM from SimPenPhen we can check the

effect of a various rates of penetrance and of phenocopy on the MapArg method. Figure 3.1 displays results of simulating data with various levels of penetrance whiles the level of phenocopy remains at 0. A more detailed explanation of the simulation process for obtaining results is given in section 3.4 but a brief description of the graphs from the data is as follows: each of the nine graphs display a likelihood profile for the MapArg method and the small triangle on the bottom axis shows the estimate of the TIM. The real position of the TIM is indicated by the vertical dotted line.

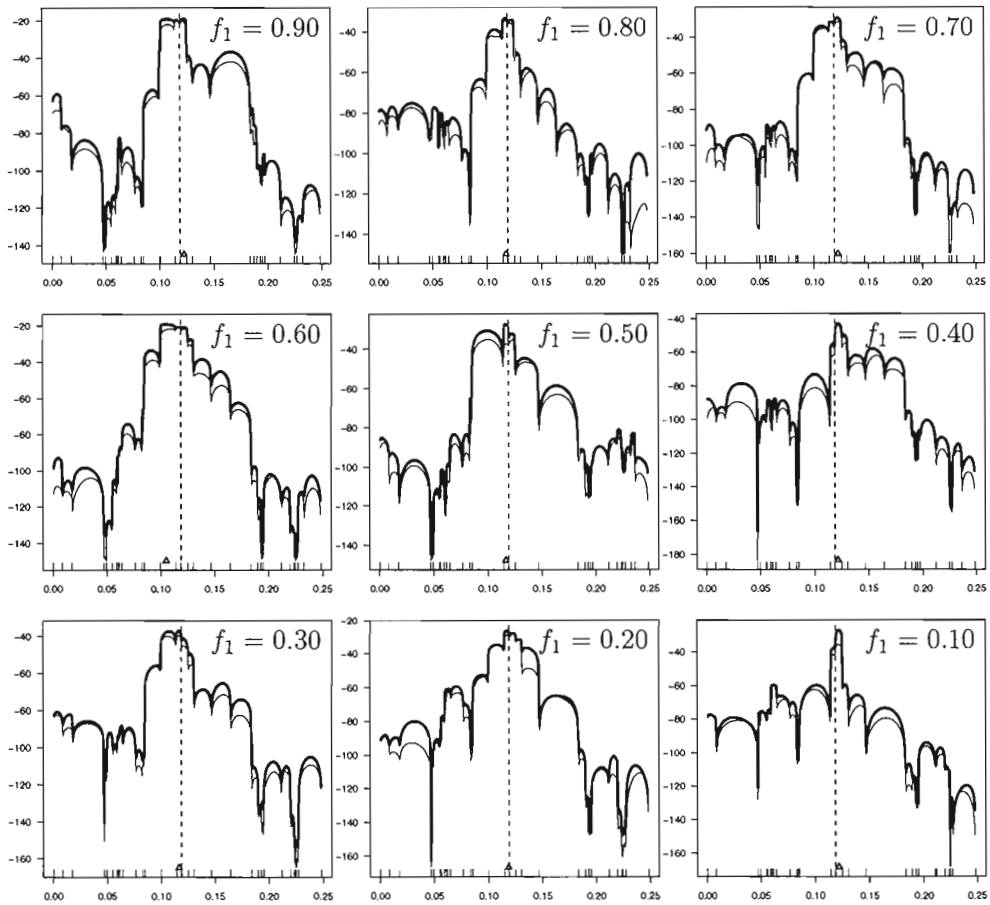

Figure 3.1 Effect of penetrance on the likelihood profile (Data B, $K = 100$, $d = 5$, $P = 30$). The level of phenocopy is 0.

In general, it seems that varying degrees of penetrance do not seem to effect the efficacy of MapArg, even when penetrance (f_1) is as low as 0.1. Imagine a population with a

disease where approximately only 10% of individuals carrying a TIM that predispose them to the disease end up being cases. Given the results shown, it appears that MapArg will still perform well in locating the position of the TIM for this population.

The existence of phenocopies, even if the rate of f_0 is low, greatly effects the ability of MapArg to find an accurate estimate for location of the TIM, as can be seen in figure 3.2 below. Notice that even for $f_0 = 0.2$, the estimate provided by MapArg is quite far from the real position of the TIM.

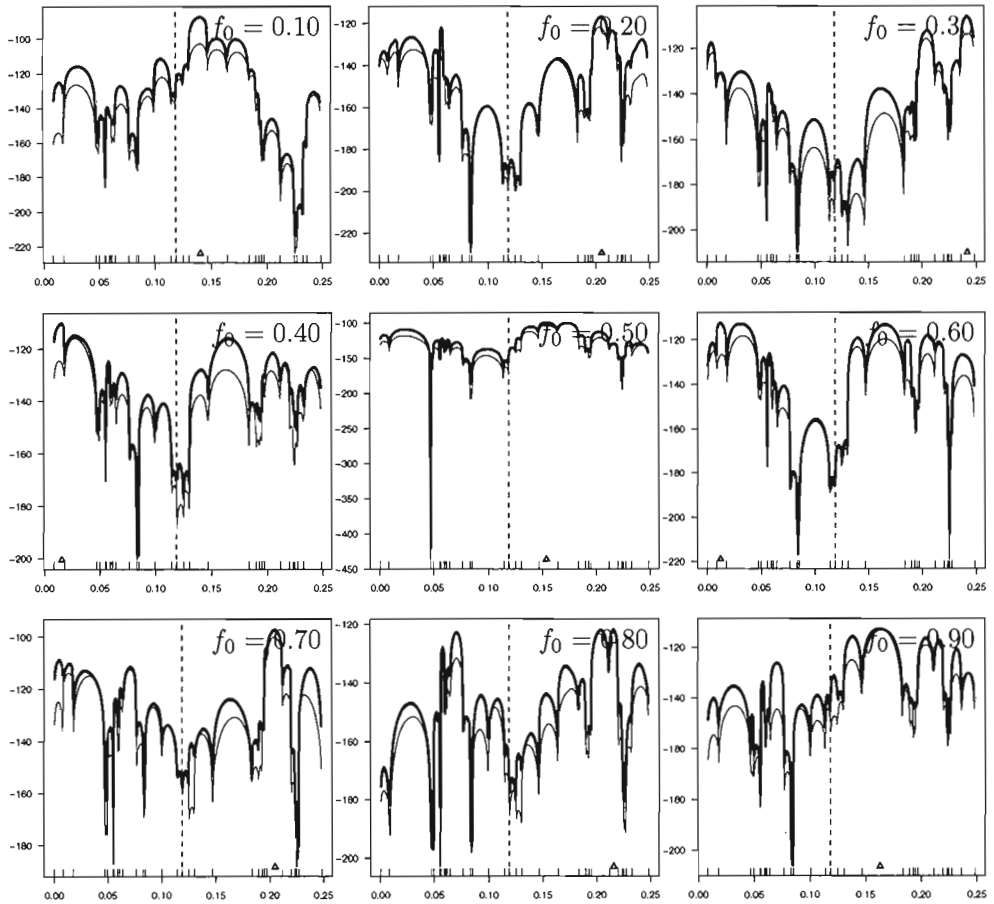

Figure 3.2 Effect of phenocopy on the likelihood profile (Data B, $K = 100$, $d = 5$, $P = 30$). The penetrance level is 1.

Intuitively, these results are not altogether surprising. Take the hypothetical situation

where there are one hundred cases and one hundred controls; 10 of the cases are phenocopies and there are 10 of the controls that have the TIM, *i.e.* $f_1 = 0.9$ and $f_0 = 0.1$. Among the population at large the genetic code of individuals is heterogeneous while the genetic coding among individuals with the TIM is similar (at least in close proximity to the TIM). Therefore if 10 individuals who have similar genetic coding (those with the TIM) are distributed among controls it should not be too noticeable as differences across the 100 controls are expected anyway. However, if among the cases, 90 have similar genetic coding but 10 differ greatly from the rest (those who don't have the TIM), the hypotheses of similarity of cases is no longer true. The model of MapArg, and indeed any fine mapping method based on LD, relies on homogeneity of cases to approximate trees and in turn estimate the location of the TIM. With phenocopies among the cases it is no longer clear if cases have differences due to recombination, which is taken into account in the model, or simply because the case does not possess the TIM in the first place. It is important therefore to develop a method that detects cases among the sample that are most likely to be phenocopies and simultaneously, controls that are likely to possess the TIM before MapArg estimates the genealogies. Section 3.3 describes two different methods that are suggested and their efficiency is tested.

3.2 Correcting For Penetrance And Phenocopy

Before going into the details of how we account for penetrance and phenocopy, it is necessary to show how the estimation of these parameters is incorporated into the methods of MapArg. Section 2.2 explains that r_T (the position of the TIM) is estimated from the distribution $Q(H_0)$, where H_0 is the haplotype state for a given sample of sequences at the present generation, for us this is the haplotype with mutation status. What we are interested in is the state for the same sample at the phenotype level which we will denote as H_{-1}. In short, we are looking for $Q(H_{-1})$, or to be more precise $Q(H_{-1}|r_T)$. Throughout this section, all calculations are conditional on r_T and so to

facilitate ease of notation we will drop r_T from subsequent equations. We can write:

$$Q(H_{-1}) = \sum_{H_0} Q(H_{-1}|H_0)Q(H_0).$$

The distribution of $Q(H_{-1}|H_0)$ does not depend on any genetic material, it depends only on f_1 and f_0. Suppose we have a sample of n sequences with n_{cases} and $n_{controls}$; conditioning on H_0, we know the number of mutants, n_m, and non-mutants, $n_{\bar{m}}$, so it is possible to evaluate:

$$Q(H_{-1}|H_0).$$

Now, let us introduce a proposal distribution $P(H_0|H_{-1})$, which will be evaluated in the subsequent section:

$$\begin{aligned} Q(H_{-1}) &= \sum_{H_0} \frac{Q(H_{-1}|H_0)}{P(H_0|H_{-1})} P(H_0|H_{-1})Q(H_0) \\ &= \sum_{H_0} f(H_{-1}, H_0)P(H_0|H_{-1})Q(H_0), \end{aligned}$$

where $f(H_{-1}, H_0) = Q(H_{-1}|H_0)/P(H_0|H_{-1})$. This gives:

$$\begin{aligned} Q(H_{-1}) &= \sum_{H_0} f(H_{-1}, H_0)P(H_0|H_{-1}) \sum_{H_1} f(H_0, H_1)P(H_1|H_0)Q(H_1) \\ &= \sum_{H_0} f(H_{-1}, H_0)P(H_0|H_{-1}) \sum_{H_1} f(H_0, H_1)P(H_1|H_0) \sum_{H_2} f(H_1, H_2)P(H_2|H_1)Q(H_2) \\ &= \ldots \\ &= \sum_{H_0} \sum_{H_1} \sum_{H_2} \ldots \sum_{H_{\tau^*}} f(H_{-1}, H_0)f(H_0, H_1)f(H_1, H_2)\ldots f(H_{\tau^*-1}, H_{\tau^*}) \\ &\quad P(H_0|H_{-1})P(H_1|H_0)P(H_2|H_1)\ldots Q(H_{\tau^*}), \end{aligned}$$

and this represents the likelihood function that estimates r_T, since:

$$L(r_T) = Q(H_{-1}) = E_P \left[\prod_{\tau=-1}^{\tau^*-1} f(H_\tau, H_{\tau+1}) \right]. \tag{3.3}$$

The genealogies are then constructed according to the distribution P. In the above expression the only terms that are unknown are $P(H_0|H_{-1})$ and $Q(H_{-1}|H_0)$, the rest can be deduced from the coalescent process. To summarize, we are able to take penetrance and phenocopy into account by modeling the sample state at H_{-1} (the phenotype level), conditioning on H_0 (the genotype level). In order to do this we will evaluate

- $Q(H_{-1}|H_0)$,
- $P(H_0|H_{-1})$.

In fact, expression (3.3) is just a generalization of the likelihood of r_T, as seen in section (2.2). If there is a perfect relationship (*i.e.* a mathematical one), between the genotypes and phenotypes $P(H_0|H_{-1})$ and $Q(H_{-1}|H_0)$ both equal 1. Then there is only one state H_0 that corresponds to H_{-1} and expression (3.3) is expression (2.2) exactly.

Evaluating $Q(H_{-1}|H_0)$

The probability that an individual i is a case or a control does not depend on the genetic code apart from the TIM, of course. Thus, we are looking at the probability of an individual being either sick (a case) or not sick (a control), given that we know if they have the TIM or not. So, $Q(H_{-1}|H_0)$ depends on the penetrance parameters uniquely. Suppose there is a certain number of cases and controls, denoted n_c and $n_{\bar{c}}$ respectively, among a sample of n sequences: $n_c + n_{\bar{c}} = n$.

For one sequence j which is a mutant ($M_j = 1$), we want to find

$$Pr(S_j = 1|M_j = 1) \text{ and } Pr(S_j = 0|M_j = 1),$$

where $S_j = 1$ if a case, 0 if a control. Similarly, for a sequence k, that is non-mutant we want

$$Pr(S_k = 1|M_k = 0) \text{ and } Pr(S_k = 0|M_j = 0).$$

These four probabilities are evaluated in section 3.3 below. For now, let us assume that they are known. The sample consists of n unrelated individuals and so putting mutant and non-mutant sequences together gives:

$$Q(H_{-1}|H_0) = \prod_{j=1}^{n_m} Pr(S_j|M_j = 1). \prod_{k=1}^{n_{\bar{m}}} Pr(S_k|M_k = 0).$$

If there are equal numbers of cases and controls in the sample, $n_m = n_{\bar{m}} = \frac{n}{2}$ and the notation simplifies to:

$$Q(H_{-1}|H_0) \quad = \quad \prod_{i=1}^{n} Pr(S_i|M_i) \quad \forall i \in n, S_i = 0, 1 \text{ and } M_i = 0, 1.$$

3.3 Evaluating $P(H_0|H_{-1})$

3.3.1 Method 1

Suppose that we know ξ, the frequency of the mutation for a population of haploids, and the rate of penetrance and phenocopy, f_1 and f_0, respectively. This simple model can be considered as an approximation of a genetic model for a recessive disease where the disease is rare (the latter representing a diploid population). Let t represent the allele responsible for the disease, *i.e.* the TIM, and T the wild type allele. Let S represent the status of an individual who is sick, and $\bar{S}$ the status of and individual who is not sick. Note that:

$$P(S) = P(S|t)P(t) + P(S|T)P(T) = f_1.\xi + f_0.(1-\xi).$$

Thus:

$$\begin{aligned} P(t|S) &= \frac{P(t \cap S)}{P(S)} = \frac{P(S|T)P(T)}{P(S)} \\ &= \frac{f_1.\xi}{f_1.\xi + f_0.(1-\xi)}. \end{aligned}$$

Similarly, we have:

$$\begin{aligned} P(t|\bar{S}) &= \frac{P(t \cap \bar{S})}{P(\bar{S})} = \frac{P(\bar{S}|T)P(T)}{P(\bar{S})} \\ &= \frac{(1-f_1).\xi}{1-[f_1.\xi + f_0(1-\xi)]}, \end{aligned}$$

and $P(T|S)$ and $P(T|\bar{S})$ can then be deduced since $P(T|S) = 1 - P(t|S)$ and $P(T|\bar{S}) = 1 - P(t|\bar{S})$. We have,

$$P(T|S) = \frac{f_0.(1-\xi)}{f_1.\xi + f_0.(1-\xi)},$$

and

$$P(T|\bar{S}) = \frac{(1-f_0).(1-\xi)}{f_1.\xi + f_0.(1-\xi)}.$$

Now, we almost have a probability distribution for $P(H_0|H_{-1})$. Sampling is not random however, so we have to make some modifications to account for this. Let $P'()$ denote the

distribution of the sample, and let $P()$ denote the population distribution. Then $P'()$ represents the proportion of cases in the sample, and $P'(\bar{S}) = 1 - P'(S)$ the proportion of controls.

Since the sample selection is a function of the case/control status we have:

$$P'(T|S) = P(T|S),$$
$$P'(T|\bar{S}) = P(T|\bar{S}).$$

Also, we have the following:

$$P'(t) = P'(t|S)P'(S) + P'(t|\bar{S})P'(\bar{S}).$$

And finally,

$$P'(S|t) = \frac{P'(t|S)P'(S)}{P'(t)}, \qquad P'(S|T) = \frac{P'(T|S)P'(S)}{P'(T)}$$
$$P'(\bar{S}|t) = 1 - P'(S|t), \qquad P'(\bar{S}|T) = 1 - P'(S|T)$$

giving a probability distribution for $P(H_0|H_{-1})$.

3.3.2 Method 2

Suppose that we know the rates of penetrance and phenocopy, f_1 and f_0 respectively, for a population of haploids. This time we will look at the relationship that exists between the penetrance parameters, f_1 and f_0, the number of cases versus controls, N_S versus $N_{\bar{S}}$, and the number of mutants versus non-mutants, N_t versus N_T in the population. If we know the number of mutants and non-mutants and the penetrance parameters in a given sample, N_S and $N_{\bar{S}}$ can be determined heuristically:

$$\begin{aligned} N_S &= N_t.f_1 + N_T.f_0 \\ N_{\bar{S}} &= N_t.(1 - f_1) + N_T.(1 - f_0) \end{aligned}$$

Note also that

$$N_S + N_{\bar{S}} = N = N_t + N_T,$$

where N is the total population size. So N_t and N_T can be expressed in terms of N_S, $N_{\bar{S}}$ and the penetrance parameters. Indeed

$$\begin{aligned} N_S &= (N - N_T).f_1 + N_t.f_0 \\ \Rightarrow N_T &= \frac{N_S - N.f_1}{f_0 - f1}, \end{aligned}$$

and we also have

$$N_t = N - N_T.$$

Once values for N_t and N_T are obtained we can then find the probability of a sequence being mutant or non-mutant.

$$P(t) = \frac{N_t}{N_t + N_t}, \quad P(T) = \frac{N_T}{N_t + N_t}.$$

Denoting the sample distribution $P^{'}()$ as in Section 1.3.1, we again make use of the fact that the sample selection depends directly on the case/control status and we have:

$$P^{'}(t) = \frac{n_t}{n_t + n_T}, \quad P^{'}(T) = \frac{n_T}{n_t + n_T}, \tag{3.4}$$

where n_t, n_T and n represent the sample number of mutant, non-mutant and total number of sequences respectively.

Let's see an example of this procedure. Consider we have 100 mutant haplotypes, and 10 non-mutant haplotypes; if f_0 is 0.15 and f_1 is 0.9, the expected number of cases and controls are:

$$100 \times f_1 + 10 \times f_0 = 91.5,$$

$$100 \times (1 - f_1) + 10 \times (1 - f_0) = 18.5,$$

respectively.

So, as we have see in formula (3.4), if we know the penetrance and phenocopy parameters and the numbers of cases and controls, one can infer the number of mutants and non-mutants. In this particular case, using formula (3.4) gives of course, 100 mutants and

10 non-mutants. We can use these estimates to build a proposal distribution to infer the mutant status that we need in H_0. But this distribution would be a bit naive.

We can do better by using the distribution of haplotypes among cases and controls, using the same simple reasoning as above, but conditional on the haplotype. As there should be linkage disequilibrium in the data, the distributions of haplotypes among cases and controls is informative. Assume we have n haplotypes of d different types ($d \leq n$). Denote $h_j^{m_1:m_2}$ ($j = 1, \ldots, d$) the partial haplotype j from marker m_1 to marker m_2; if $m_1 = 1$ and $m_2 = L$, then $h_j^{m_1:m_2}$ is just the whole haplotype of an individual of type j. Moreover, denote n_j^c and $n_j^{\bar{c}}$ the number of cases and controls among the haplotypes of type j, such as $n_j^c + n_j^{\bar{c}} = n_j$. Then, n_j^t, an estimate of the number of mutants haplotypes among the haplotypes of type j can be estimated; similarly, n_j^T can also be estimated. Using the same reasoning as above, we have for $j = 1, ..., d$:

$$\begin{aligned} P[M_i = 1 \mid \text{seq } i \text{ is of type } j] &= \frac{n_j^t}{n_j^t + n_j^T}, \\ P[M_i = 0 \mid \text{seq } i \text{ is of type } j] &= \frac{n_j^T}{n_j^t + n_j^T}, \end{aligned}$$

where

$$n_i^T = \frac{n_i^c - n_i \cdot f_1}{f_0 - f_1}, \quad n_i^t = n_i - n_i^T.$$

As we have seen in chapter 2, MapArg builds graphs interval by interval: this means that the graphs are generated differently depending on the candidate value of r_T at which MapArg is currently evaluated. Near the real position of the TIM, mutant haplotypes should differ more from non mutant haplotypes than anywhere else in the sequence, hence the proposal distribution should be more accurate around the real position of the TIM. At a given position, the length of the haplotypes used to build the proposal distribution is $m_2 - m_1$. We can expect the length to have an impact on the quality

of the proposal distribution: too short will not bring enough information, but too long would just add noise.

3.4 Simulations

Simulating Sample Data Sets

To implement MapArg whiles correcting for penetrance and phenocopy, we first need to simulate a population of sequences and in turn sample a data set that resembles a real population of individuals with a TIM among them. The program *ms* of Hudson (2002) is used to generate the sequences. This program generates sequences under recombination in a neutral population of constant size. Here 1 sample of 10 000 sequences is generated. This sample is generated for a fixed value of the scaled recombination rate that corresponds to a fixed number of sites (loci) along the whole sequence. The scaled recombination rate chosen is $\rho = 100$ for 0.25Mb, and the number of sites chosen is 80. Even though there are a finite number of sites, mutations are assumed to occur according to the infinite sites model (See Section 2.2.1), therefore it is assumed that the mutation occurs once and once only among the population. The scaled recombination rate ρ_m between markers m and $m+1$, is converted to a genetic scale, using $r_m = \rho_m/(4N)$, where $N = 10\ 0000$ is the constant population size. Now we have an approximation for the distances between markers on the cM scale (See Section 1.2.1 for an explanation of the Morgan scale).

Recall from Section 1.1.3 that polymorphisms are common differences in the sequence of DNA, occurring in at least 1% of the population. So the more polymorphic a site is, the higher the probability of the less common allele appearing at this loci is. If sites are not very polymorphic then we know that the less common allele has a low probability of appearing, resulting in most sequences carrying the more common allele (wild type) and so sequences are indistinguishable from one another. The most polymorphic sites are then chosen.

The position of the TIM is chosen such that it is located between the first and the third quartiles of the sequence, to ensure there are markers on both sides of the TIM. As before, ξ is the frequency of the mutation among the population of sequences and it is set at 0.1. The sample has a sequence of 0.25cM in length and with a mutation rate of 0.1, which corresponds to a sample from a population with a common disease. Sequences that carry the mutation at the locus for the TIM are considered cases whiles those who don't are considered to be controls.

Estimating the position of the TIM is done using a subsample of the larger sample, which is assumed to be representative of the population. Ideally a random sample would be selected. However, when the disease frequency is low the subsample would have to be extremely large if we want to ensure a significant number of cases in our subsample. For this reason it is necessary to fix the number of cases and controls so that a minimum of information for both groups is obtained. For our analysis, the subsample generated called Data B, consists of 100 cases and 100 controls that have been drawn at random without replacement from their respective samples.

Incorporating $P(H_0|H_1)$ into the MapArg Framework

We have described two different methods that take incomplete penetrance and phenocopy into account for the fine mapping method MapArg. Also, it has been deduced by means of simulation that incomplete penetrance does not have too much of an effect on MapArg's ability to estimate r_T, whereas even small levels of phenocopy can have a negative effect on its performance. In order to see the effects of Method 1 and Method 2 above (See section 3.3), we need only extend the C++ program, already in existence for MapArg to incorporate each method in turn. Several simulations are run for each method in turn, with various values for f_1 and f_0. Results of these simulations are presented and discussed in the following section.

3.5 Results In MapArg Accounting For The Penetrance Paramaters

3.5.1 A Description Of The Graphs in MapArg

All the analyses have been done with the composite likelihood strategy, with windows of length 5 markers, and using 30 markers (or otherwise specified). Every analysis has been carried out on the data set Data B (See Section 3.4.1) apart from those of Section 3.4.3, where the Cystic Fibrosis data is analyzed. The data set B is known to behave as expected in the LD theory, so is very useful when exploring new strategies of analysis. Figure 3.3 shows the "real" likelihood of the data Data B, *i.e.* the sample of 50 cases and 50 controls from the population B, with no phenocopy and full penetrance ($f_0 = 0$, $f_1 = 1$). The x axis is the location of the the TIM in the sequence, and the y axis is the logarithm of the likelihood. The estimate $L(\hat{r_T})$ is indicated by the triangle at the bottom of the figure, and the real position of the TIM by the vertical dotted line. As we can see on figure 3.3, in this case, the estimate is just near the real location of the

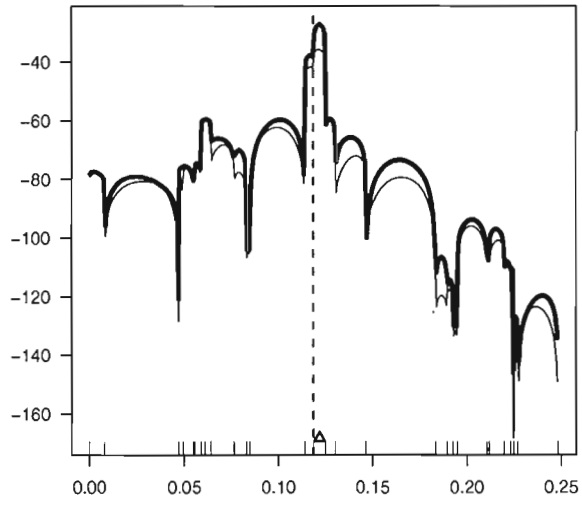

Figure 3.3 The likelihood profile for data B, with windows of 5 markers at a time and a total of 30 markers. Here $f_0 = 0$ and $f_1 = 1$.

TIM. In the simulations which follows, we should ideally see a similar profile to this one, whatever the penetrance and phenocopy are. The little bars at the bottom of the graph indicates the position of the markers.

3.5.2 Results On Simulated Data

First, we take a look at some output for MapArg when it corrects for the penetrance parameters with Method 1 (See Section 3.3.1). Selected results are presented here to see the efficacy of Method when levels of penetrance decrease and levels of phenocopy increase. Results for Method 2 are then presented. It is hoped that the performance of MapArg is better when adjusting for each method in turn, at least for under some conditions (e.g. low levels of phenocopy, large sample size etc.).

MapArg with Method 1

It seems from figure 3.4 that MapArg remains very efficient for varying levels of penetrance. However, we have already seen that penetrance does not seem to effect the efficacy of MapArg, even when there is no method modeling incomplete penetrance. When there is a low level of phenocopy ($f_0 = 0.1$), Method 1 works quite well. However for $f_0 = 0.2$ or greater, the method is not too successful as can be seen in figure 3.5. This is a reflection of the fact that MapArg has difficult finding the location of the TIM when there are phenocopies in the data. We have combined incomplete penetrance and phenocopy in Figure 3.6 and it appears that Method 1 is no longer capable of improving the performance of MapArg. This is most likely due to the very strong effect that phenocopy has on the efficiency of MapArg.

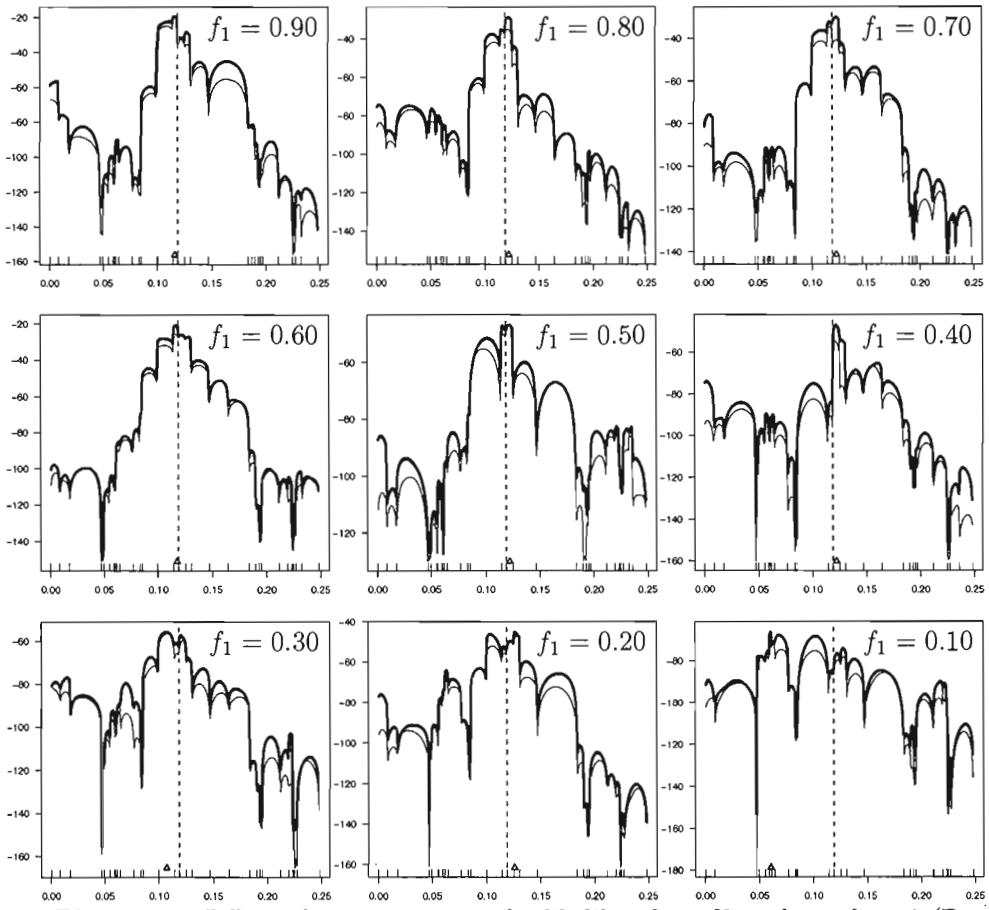

Figure 3.4 Effect of penetrance on the likelihood profile, where $f_0 = 0$ (Data B, K = 10, d = 5, P = 30).

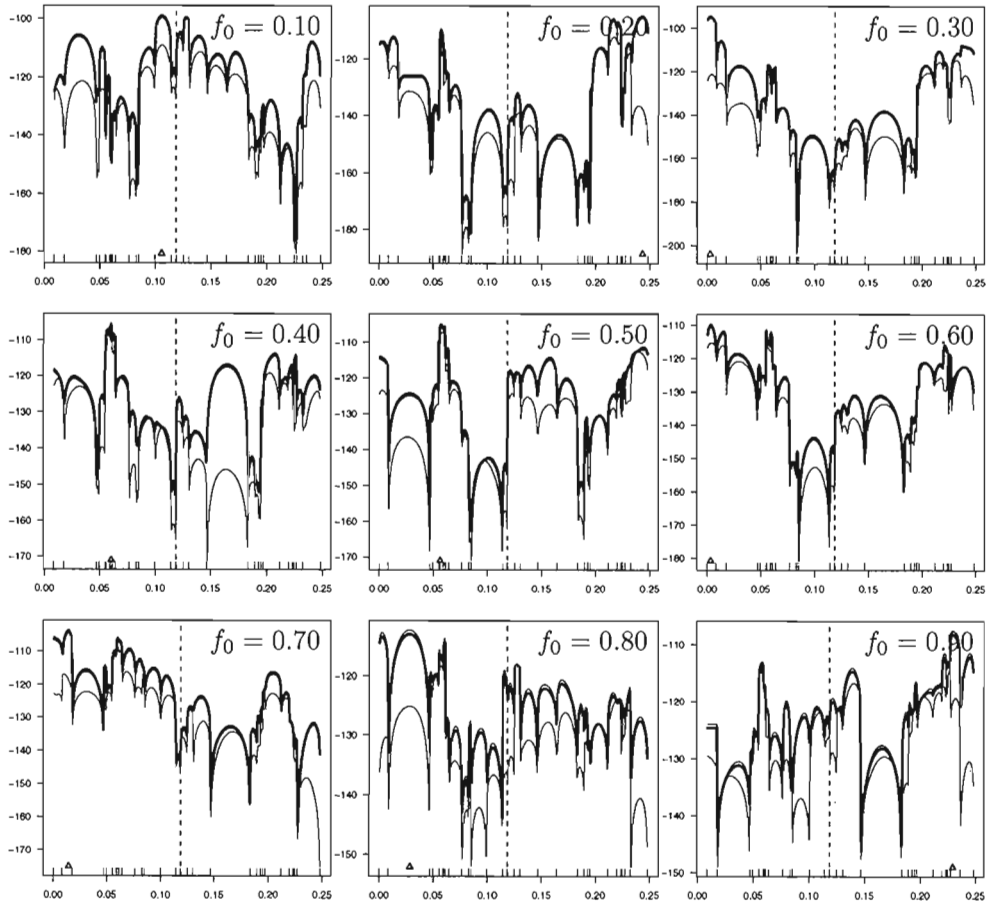

Figure 3.5 Effect of phenocopy on the likelihood profile, where $f_1 = 1$ (Data B, K = 10, d = 5, P = 30).

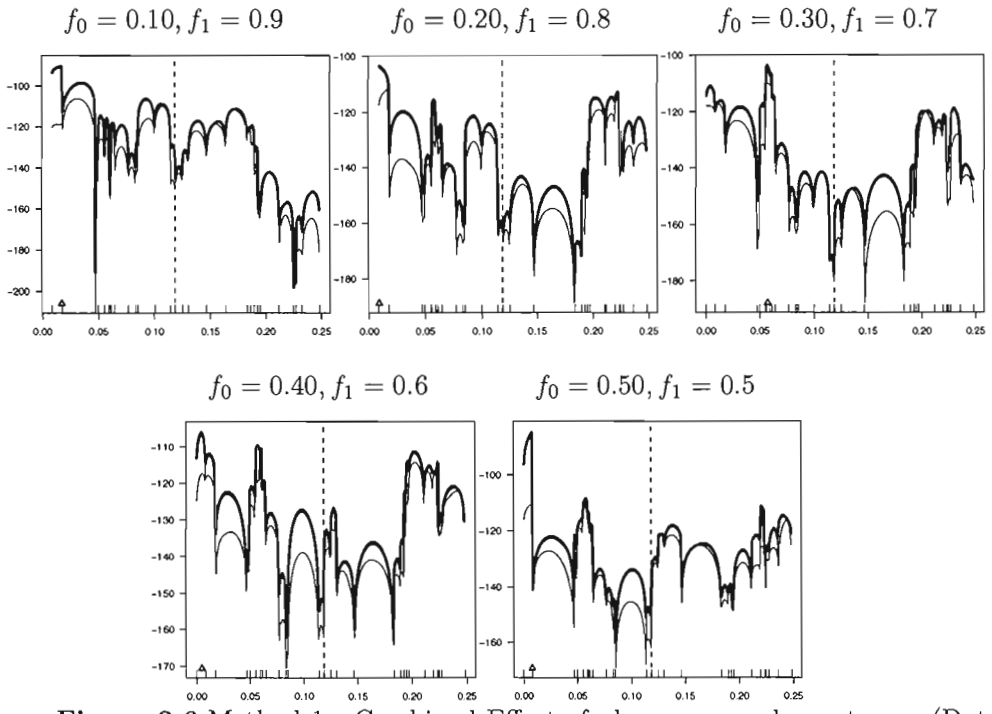

Figure 3.6 Method 1 - Combined Effect of phenocopy and penetrance (Data B, K = 10, d = 5, P = 30).

MapArg with Method 2

In the following section of results we present the effects of penetrance and phenocopy separately. In each graph the left column represents the likelihood profile with incorporating Method 2 into the MapArg framework. The middle and right columns are the likelihoods with Method 2 and with partial haplotype used to build the proposal distribution, of length 4 and 8 respectively. In fact l= 2 and l= 4 in the graph represent the number of makers each side of the TIM that are used and henceforth we shall denote l as the half-window length (not to be confused with d the window length in the composite likelihood of MapArg).

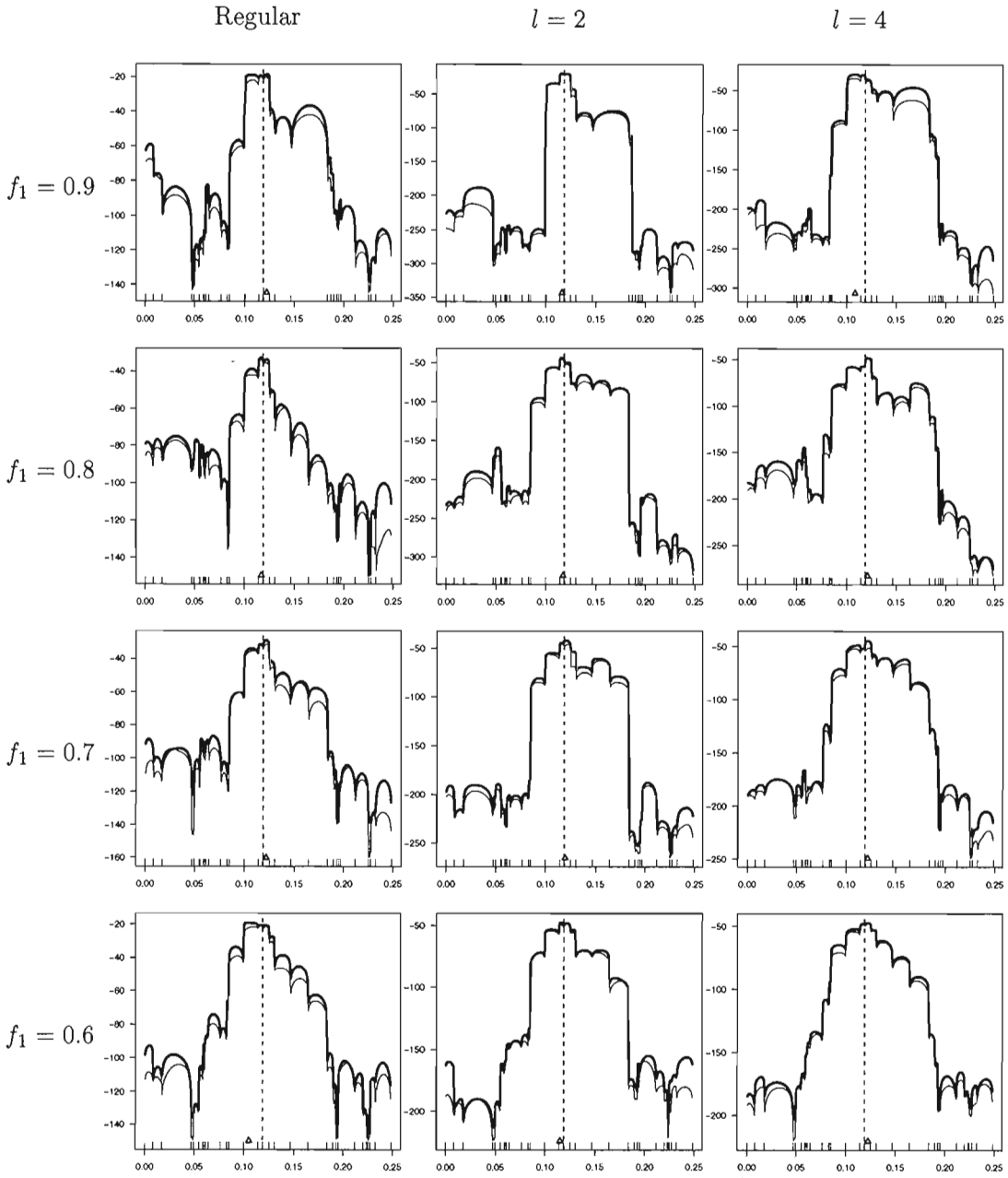

Figure 3.7 Effect of penetrance on the likelihood profile (Data B, $K = 1t$, $L = 5$, $P = 30$).

As MapArg is not affected much by incomplete penetrance we present results for f_1 with values between 0.9 and 0.6 only. Method 2 is efficient both when $l = 2$ and when $l = 4$, as is MapArg without any adjustment for the method. Results are displayed in

figure 3.7. When there is a low level of phenocopy, Method 2 performs adequately for each half-window length. However for higher levels it becomes less and less efficient, unfortunately, as can be seen in figure 3.8. When penetrance and phenocopy both exist

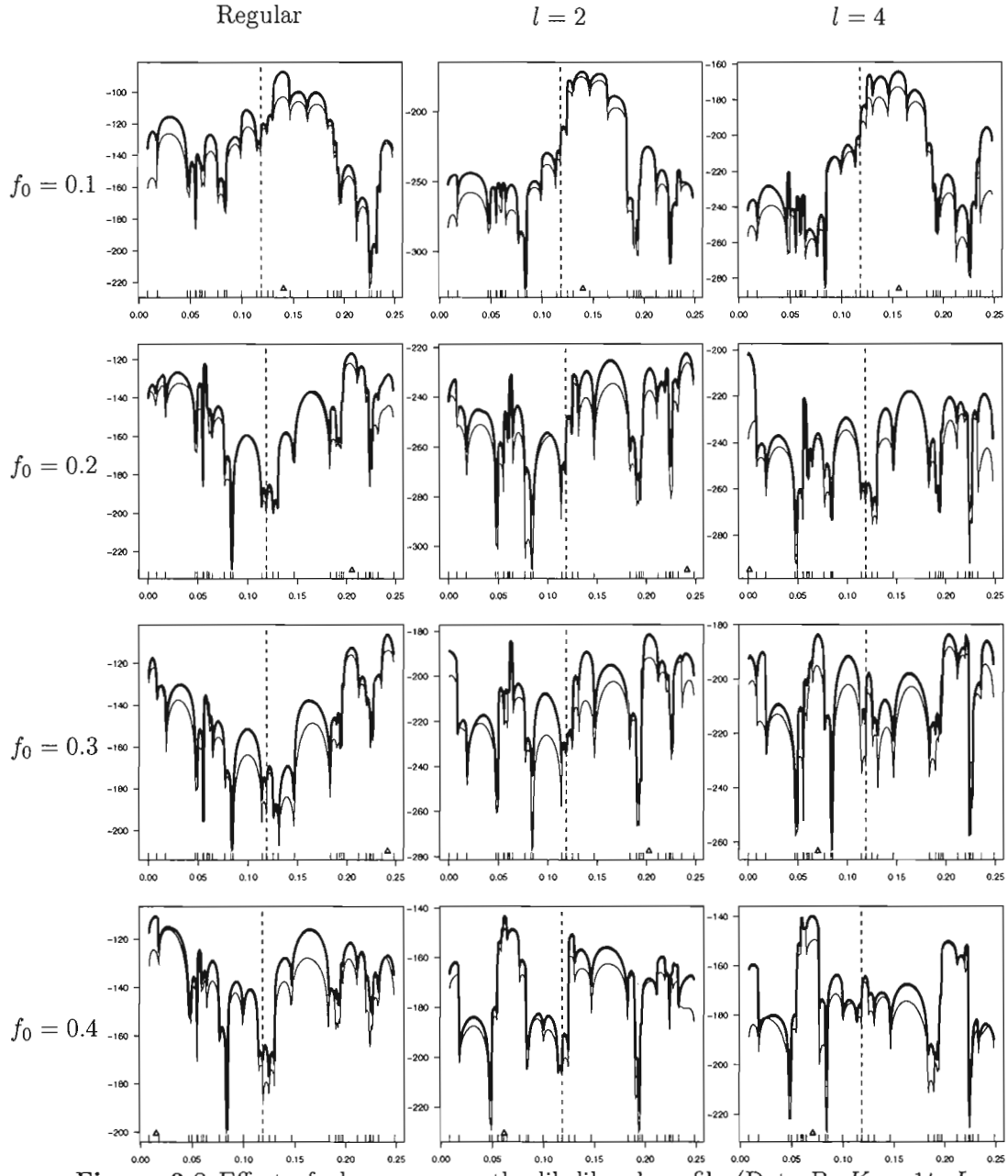

Figure 3.8 Effect of phenocopy on the likelihood profile (Data B, $K = 1t$, $L = 5$, $P = 30$).

the conclusion is not any different (Results not shown here).

3.5.3 Effect of sample size

It is of interest to see if Methods 1 or 2 prove more efficient for larger sample sizes. To see the effect of sample size, we have arbitrarily chosen a given set of parameters, $f_0 = 0.4$ and $f_1 = 1$. We have seen in Section 3.1 that with a level of phenocopy this high it is difficult to locate the TIM. This time the left column in the graph represents Method 1 and the middle and right columns represent Method 2 with different half-window lengths. Each row represents an increase in the sample size. As we can see, from Figure 3.9, Method 1 does not improve with increasing sample sizes. This is of course expected since this method does not use any information on the sample. We clearly see that increasing the sample size greatly improves the efficiency of Method 2, for both values of l. This is a very positive result as it means that if a population with a disease that is expected to have high levels of phenocopy is being analyzed, incorporating Method 2 along with a large sample size from the given population could prove very efficient.

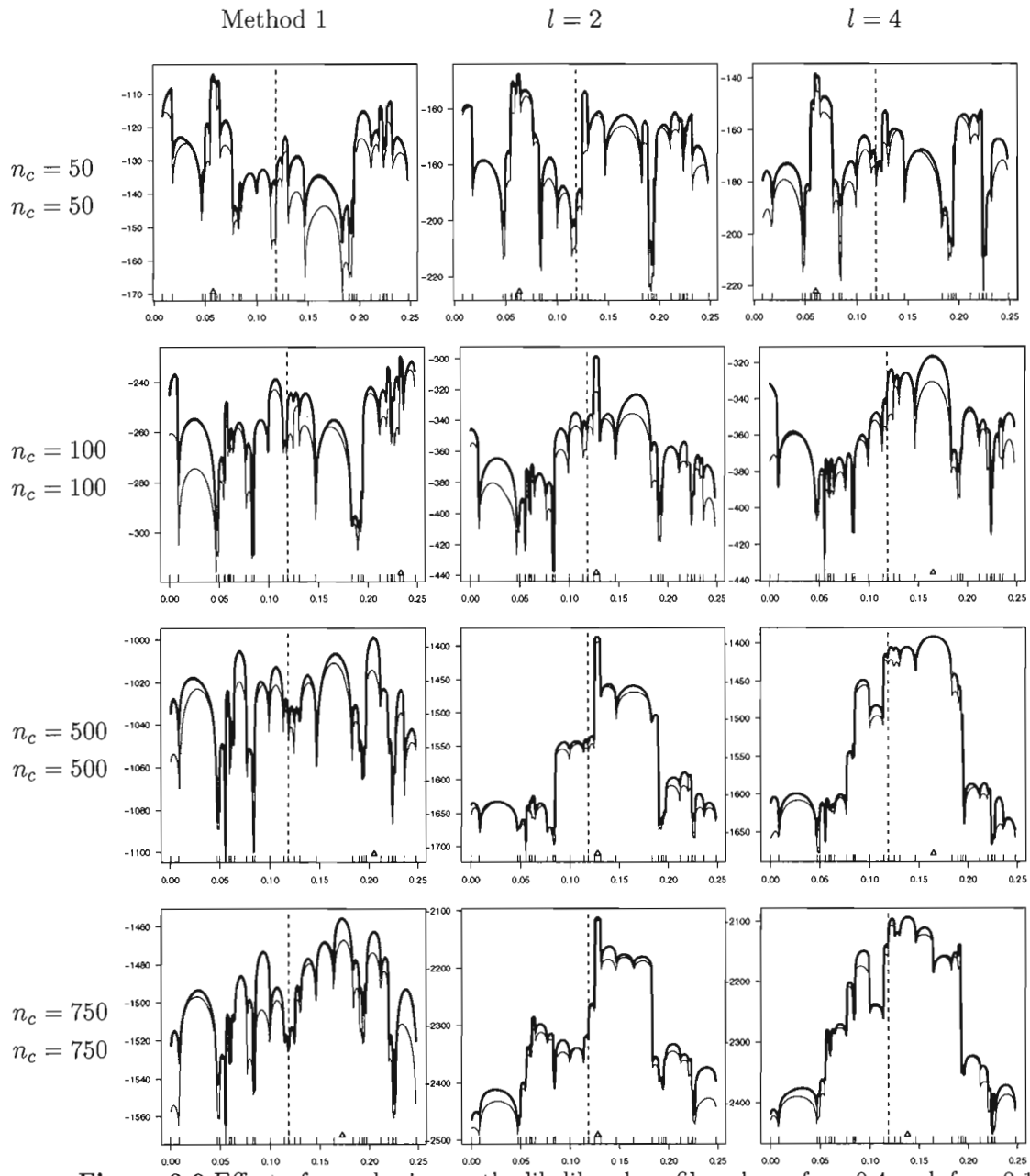

Figure 3.9 Effect of sample size on the likelihood profile, where $f_0 = 0.4$ and $f_1 = 0.1$ (Data B, $K = 1t$, $L = 5$, $P = 30$).

3.5.4 Effect of half-window length, (l)

Method 2 uses information from the the sample data to correct for the penetrance parameters by means of partial haplotypes, which are equivalent to $2l$ in our simulations.

We expect that if the length of l is too small that we would not have much of an improvement on the efficiency of MapArg. How ever if l is too large then we are using all of the sample and this too could prove inefficient. Figure 3.9 presents the effect of different half-window lengths, to give an overview of the effects of the amount information around the TIM that is used in Method 2. It seems from the graphs that efficiency is improved as l increases, but only up to a certain point. Once l is over 15, *i.e.* haplotypes of length 30 are used in Method 2, the results become less and less accurate.

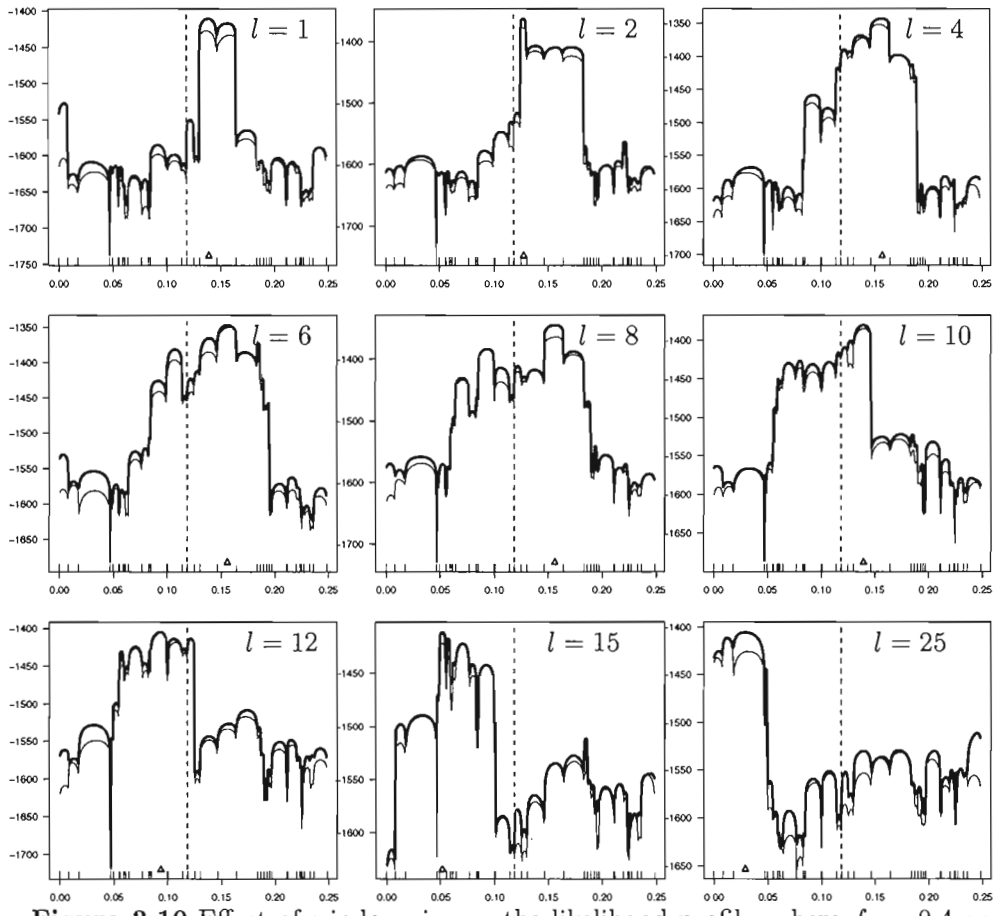

Figure 3.10 Effect of window size on the likelihood profile, where $f_0 = 0.4$ and $f_1 = 0.1$ (Data B, $K = 1t$, $L = 5$, $P = 30$).

3.5.5 Results with the Cystic Fibrosis Data Set

Cystic Fibrosis (CF) is a recessive disorder occurring in Caucasian populations and it is well known in the literature. The position of the TIM has been located on chromosome 7q31 and it is well known that ΔF508 accounts for most of the mutations in the same gene but that there are other alleles causing the disease also. The CF data have markers that are known as microsatellites which are markers that are more polymorphic than SNPs but for simplification we consider the markers as SNPs. There are 94 cases and 92 control haplotypes in the data set as given in Kerem *et al.* (1989). Let us assume for now that the 92 controls are non-mutant, so that $f_1 = 1$. Given that there are phenocopies in the data we choose a level of phenocopy, $f_0 = 0.25$, that is quite plausible as ΔF508 accounts for most but not all of the mutations that cause CF. Figure 3.11 gives results for Method 2. It seems unreasonable to assume that $f_1 = 1$, so we try a smaller penetrance of $f_{0.8}$. This gives figure 3.12

The graphs in figures 3.11 and 3.12 compare the unadjusted likelihood profile with Method 2 for half-window lengths of 2 and 4. Composite likelihood is used as in every other simulation but these time varying windows lengths denoted d, for the likelihood are used. Figure 3.11 shows that when Method 2 is incorporated into the MapArg framework the results are very encouraging. Whiles $l = 2$ is slightly more efficient than when $l = 4$ the two sets of graphs show that the estimate for the TIM location is very accurate and is also quite consistent across various values for d. Figure 3.12 assumes that penetrance is not complete ($f_1 = 0.8$) and it can been seen that Method 2 is even more efficient than when assuming complete penetrance. In fact for $l = 2$, we can see that MapArg is capable of locating the TIM for several values of d. This finding is consistent with results shown in Figure 3.9. The CF data has close to 100 cases and controls, and in figure 3.9 we see that Method 2 is very good at finding the true location of the TIM.

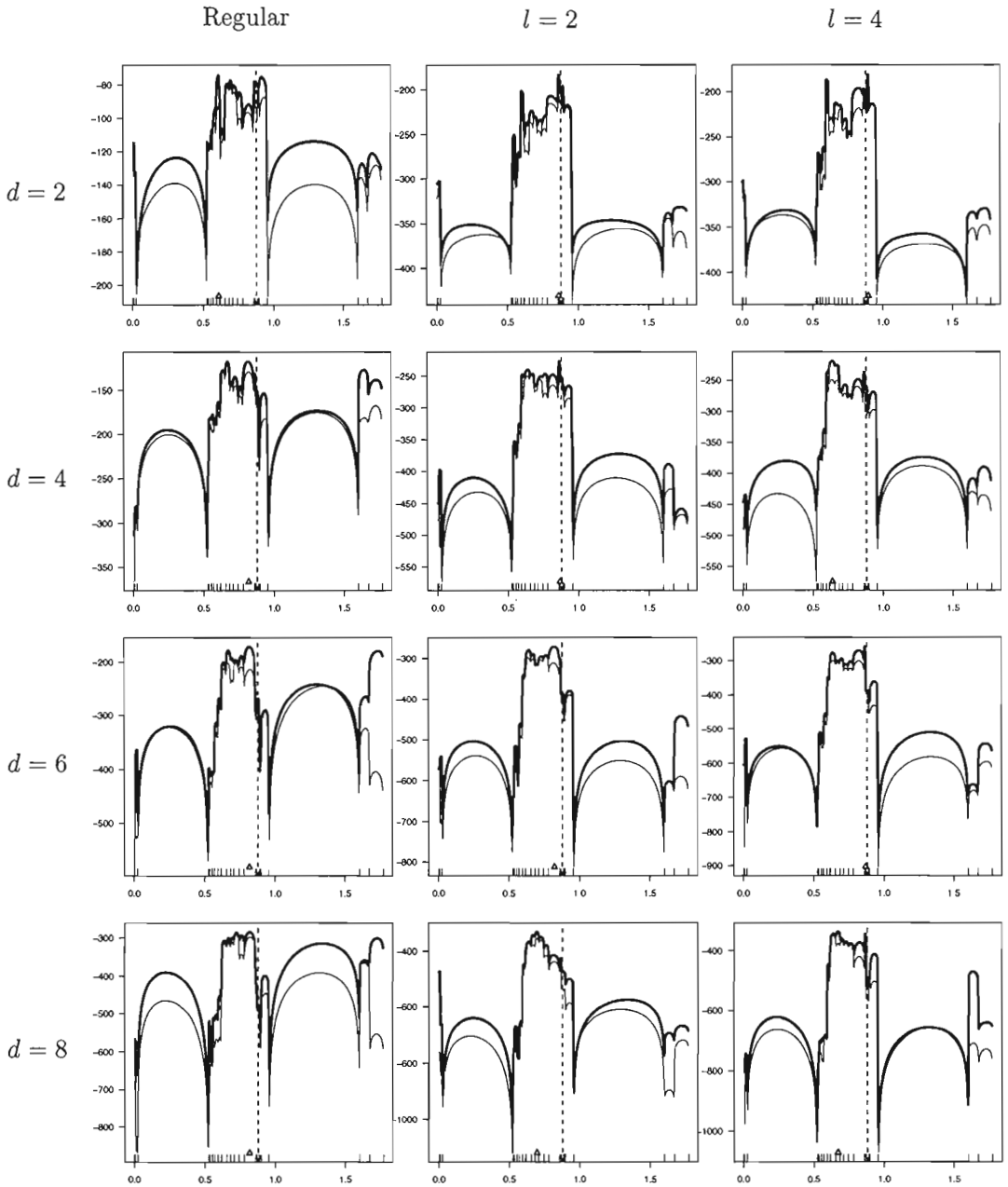

Figure 3.11 Efficiency of Method 2 with the CF data with $f_0 = 0.25$ and $f_1 = 1$.

Figure 3.13 shows the effect of Method 2 when using a window size (d), of 10. The results are still very accurate with Method 2 and are consistent with Figures 3.11 and 3.12. We present the middle graph only in Figure 3.14 and it is clear that Method 2 improves efficiency for the Cystic Fibrosis data.

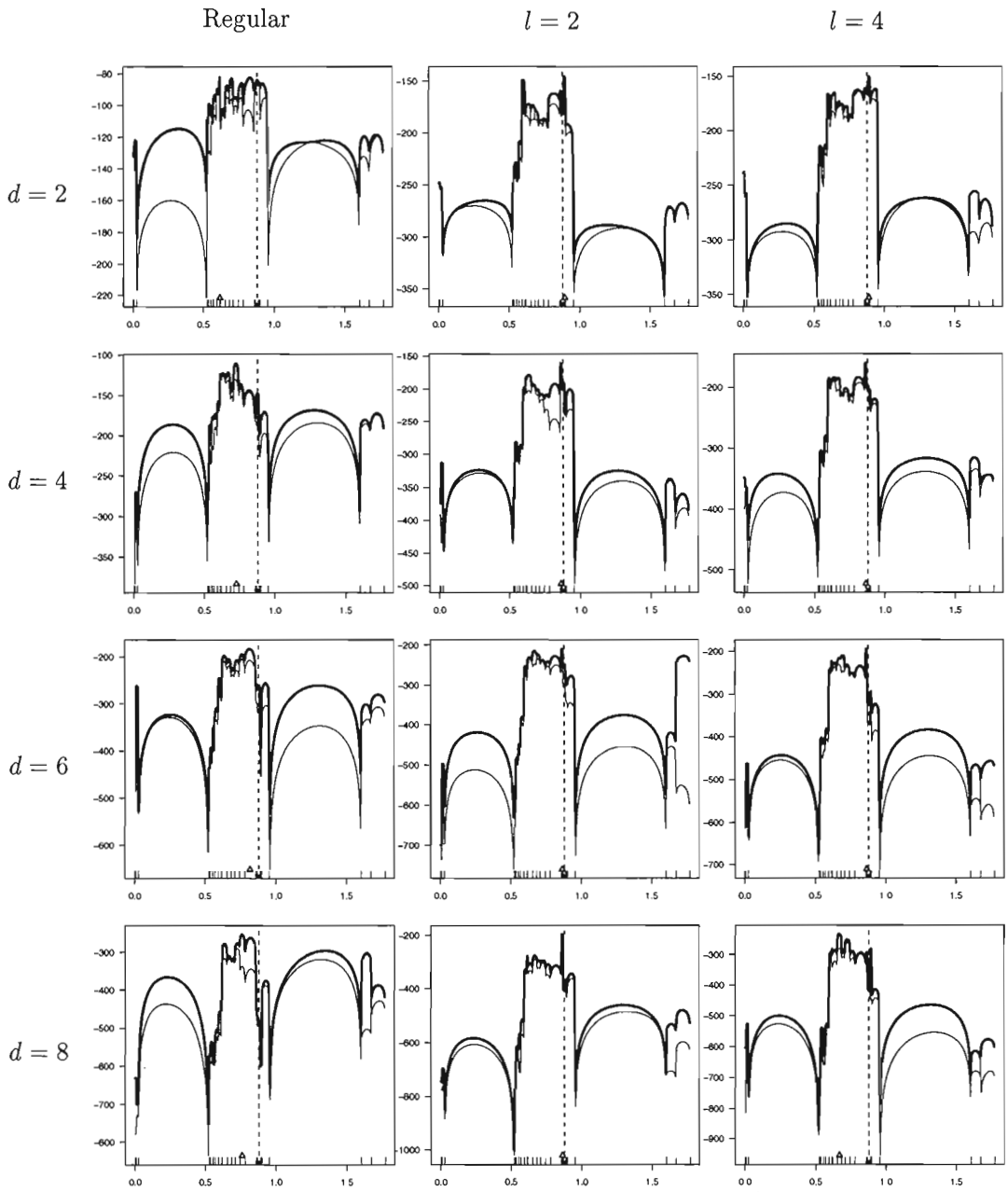

Figure3.12 Efficiency of Method 2 with the CF data with $f_0 = 0.25$ and $f_1 = 0.8$.

3.6 Further Developments

The results shown above indicate that for certain situations Method 2 improves the performance of MapArg when there is incomplete penetrance and phenocopy present.

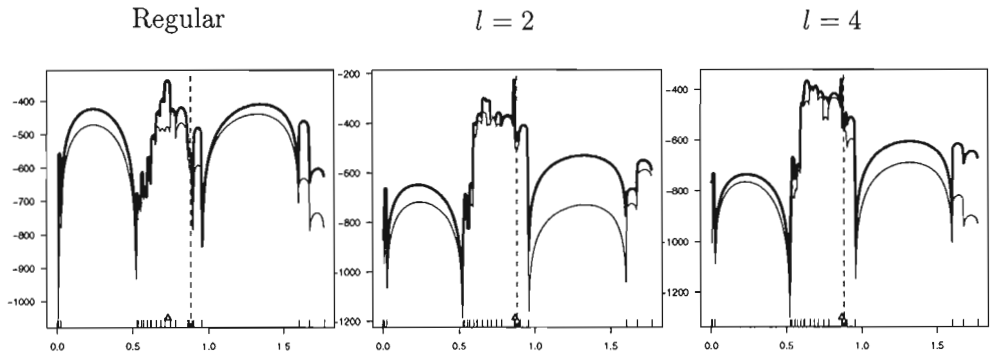

Figure 3.13 Efficiency of Method 2 with the CF data with a window of size 10 ($d = 10$).

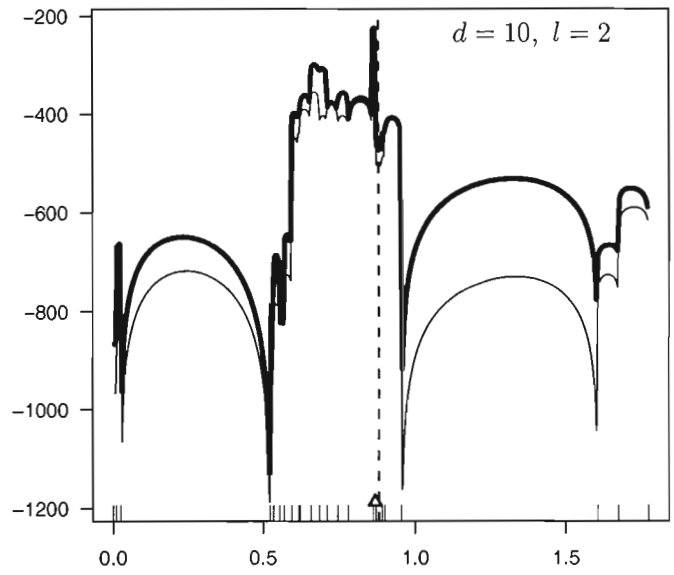

Figure 3.14 CF data with a window of size 10 ($d = 10$) and half window size of 2 ($l = 2$).

In particular, when phenocopies are present Method 2 renders MapArg more accurate at estimating r_T, for large sample sizes. Method 1 is rather inefficient but seems to be able to help in some cases. It is possible to develop other methods that correct for the penetrance parameters and it is of particular interest to develop this method for diploid data. The approach of Method 1 is rather naive but it is a good starting place for modeling the penetrance parameters, particularly when MapArg will correct for f_1 and

f_0 for diploid data. An extension of Method 1, that works with diploid data is given in the following section.

3.6.1 Method 1 with Diploids

The Model

Until now we have supposed implicitly that we are dealing with a recessive model: each individual possesses the mutation on each of their two chromosomes. Now, we suppose a more general model. Let's suppose a population of $N/2$ diploids or N haploids, in other words $N/2$ individuals, N haplotypes. As before, t represents the mutant allele at the locus of the TIM and T represents the the wild type and S_i denote the affected status of phenotype for individual i, where $S_i = 1$ for cases and 0 for controls, and $i\epsilon(1, ...N/2)$. Now the penetrance function has three parameters as opposed to two for a haploid population. The penetrance function $F = (f_0, f_1, f_2)$, is defined as:

$$\begin{cases} f_0 = P(S_i = 1|T,T) \\ f_1 = P(S_i = 1|t,T) \\ f_2 = P(S_i = 1|t,t). \end{cases}$$

A recessive model corresponds to F = (0,0,1), meaning the only possibility for a diploid population to display signs of being effected by a TIM is if both chromosomes carry the mutant allele. In fact, until now we have implicitly worked with a recessive model. A dominant model is one for which F = (0,1,1). Therefore, even in only one of two the chromosomes carry the mutant allele the diploid will be effected.

As mentioned earlier (See Section 3.2), the graphs in MapArg are produced starting at H_0, which is the haplotype data at genotype level. Define the state immediately before this one, *i.e.* the phenotype level and denote it H_{-1}:

H_1 : Haplotype with the trait (*i.e.* phenotype level),

H_0 : Haplotype with mutant status (*i.e.* genotype level).

Note that H_{-1} can also represent diploid data as well as haploid data *i.e.* H_{-1} represents the phenotype associated with whole genotype (two sequences) for each individual in diploid data as opposed to one sequence. Suppose for the moment that the haplotypes are known. We are looking for $Q(H_{-1})$, or to be more precise $Q(H_{-1}|r_T)$, where r_T represents the position of the TIM as previously defined; and everything is conditional on r_T so we drop this parameter form the notation. We can write:

$$Q(H_{-1}) = \sum_{H_0} Q(H_{-1}|H_0)Q(H_0)$$

The distribution $Q(H_{-1}|H_0)$ depends uniquely on the penetrance function F, and no information on the genotype is needed. Now, let us introduce a proposal distribution $P(H_0|H_{-1})$:

$$\begin{aligned} Q(H_{-1}) &= \sum_{H_0} \frac{Q(H_{-1}|H_0)}{P(H_0|H_{-1})} P(H_0|H_{-1})Q(H_0) \\ &= \sum_{H_0} f(H_{-1}, H_0)P(H_0|H_{-1})Q(H_0), \end{aligned}$$

where $f(H_{-1}, H_0) = Q(H_{-1}|H_0)/P(H_0|H_{-1})$ and as before, we can write the likelihood function that estimates r_T:

$$L(r_T) = Q(H_{-1}) = E_P \left[\prod_{\tau=-1}^{\tau^*-1} f(H_\tau, H_{\tau+1}) \right]. \tag{3.5}$$

Notice that the development of $Q(H_{-1})$ is analogous to that of $Q(H_{-1})$ for haploid data (See expression (3.3)). The only point in which they differ is in the estimation of the two unknown terms above, namely:

- $Q(H_{-1}|H_0)$,
- $P(H_0|H_{-1})$.

Estimation of these two terms for diploid data will soon be explained. Beforehand, it is necessary to express ξ, frequency of the mutation in the population, as a function of F.

3.6.2 The Frequency of the Mutation for Diploid Data

Suppose the frequency of the mutation among the population of $2N$ chromosomes is ξ. Furthermore, suppose the genetic model $F = (f_0, f_1, f_2)$ is known. Let n_c and $n_{\bar{c}}$ be the number of cases and controls in the sample, and f_c and $f_{\bar{c}}$, their respective frequencies. We need to estimate ξ as a function of n_c and $n_{\bar{c}}$. Since ξ represents the mutation rate at the population level, we deduce that the individuals of the population are:

$$\begin{cases} tt, & \text{with probability } (1-\xi)^2 \\ tT, & \text{with probability } 2(1-\xi).\xi \\ TT, & \text{with probability } \xi^2. \end{cases}$$

Hence, the average frequencies of cases (f_c) and controls ($f_{\bar{c}}$) are:

$$\begin{cases} f_c = (1-\xi)^2.f_0 + 2(1-\xi).\xi.f_1 + \xi^2.f_2 \\ f_{\bar{c}} = (1-\xi)^2.(1-f_0) + 2(1-\xi).\xi.(1-f_1) + \xi^2.(1-f_2). \end{cases}$$

and it is possible to solve for ξ:

Therefore we can estimate ξ as a function of $F = (f_0, f_1, f_2)$, if we have a random sample. Unfortunately, this is not the case, since the mutation is usually rare: so in order to have a representative sample we choose n_c and $n_{\bar{c}}$ to be similar in size. It is reasonable to assume however that the researcher has a good estimate of the frequency of the disease at population level, so we can still calculate the frequency of the mutation with the penetrance function F.

Evaluation of $Q(H_{-1}|H_0)$

Here, H_0 corresponds to the genetic data with information of the TIM. From a written viewpoint this is rather delicate, because we need to consider information in pairs of haplotypes within H_0. A pair is tt, tT or TT. The probability that a particular pair of exhibit the disease or not in H_1 depends only on the penetrance function F, which we consider known. Denote S_i, a random variable that equals 1 if individual i is a case and 0 otherwise, and G_i, the genotype of of the individual i, where $G_i = (tt, tT, TT)$.

For ease of notation, we will write $P(G_i = TT)$ as $P(TT)$. We have:

$$\begin{cases} P(S_i = 1|TT) = f_2 \\ P(S_i = 1|tT) = f_1 \\ P(S_i = 1|tt) = f_0. \end{cases}$$

Evaluation of $P(H_0|H_{-1})$

This proposal distribution, is the distribution that renders the method efficient if well chosen and possibly less efficient otherwise. For a given individual i, we look for the probability that there genotype is TT, Tt or tt. This probability depends on the penetrance function F:

$$\begin{aligned} P(TT|S_i = 1) &= \frac{P(TT \bigcap S_i = 1)}{P(S_(i) = 1))} \\ &= \frac{P(S_i) = 1|TTP(TT)}{P(P_i = 1)} \\ &= \frac{f_2.\xi^2}{f + c} = \frac{f_2.\xi^2}{(1-\xi)^2.f_0 + 2(1-\xi).\xi.f_1 + \xi^2.f_2} \end{aligned}$$

The three probabilities for cases are obtained similarly, and we have:

$$\begin{aligned} P(TT|S_i = 1) &= \frac{f_2.\xi^2}{(1-\xi)^2.f_0 + 2(1-\xi).\xi.f_1 + \xi^2.f_2} \\ P(tT|S_i = 1) &= \frac{f1.2(1-\xi).\xi}{(1-\xi)^2.f_0 + 2(1-\xi).\xi.f_1 + \xi^2.f_2} \\ P(TT|S_i = 1) &= \frac{f_0.(1-\xi)^2}{(1-\xi)^2.f_0 + 2(1-\xi).\xi.f_1 + \xi^2.f_2}, \end{aligned}$$

and similarly controls may be deduced. However, the sample is not randomly chosen. Let g_c be the case frequency of the sample, which is known. We are looking for $P^{'}(H_0|H_{-1})$, where $P^{'}()$ represents the proposal distribution for the sample. Note that

$$P^{'}(TT|S_i = 1) = P(TT|S_i = 1)$$

as we sample with respect to the status cases and controls. The other probabilities are obtained similarly. We now want to find:

$$P^{'}(S_i = 1|TT) = \frac{P^{'}(TT|S_i = 1)P^{'}(S_i = 1)}{P^{'}(TT)}$$

Note that:

$$\begin{aligned} P^{'}(TT) &= P^{'}(TT|S_i=1)P^{'}(S_i=1)+P^{'}(TT|S_i=0)P^{'}(S_i=0) \\ &= P(TT|S_i=1).g_c+P(TT|S_i=0).g_{\bar{c}} \\ &= \frac{f_2\xi^2}{f_c}.g_c+\frac{(1-f_2)\xi^2}{1-f_c}.(1-g_c) \end{aligned}$$

Similarly,

$$P^{'}(Tt)=\frac{f_1 2(1-\xi)\xi}{f_c}.g_c+\frac{(1-f_1)2(1-\xi)\xi}{1-f_c}.(1-g_c),$$

and

$$P^{'}(tt)=\frac{f_0(1-\xi)^2}{f_c}.g_c+\frac{(1-f_0)(1-\xi)^2}{1-f_c}.(1-g_c).$$

Therefore, we are able to calculate

$$P(G_i|S_i), \text{and} P(S_i|G_i).$$

We use $P(G_i|S_i)$ as the proposal distribution ($P(H_0|H_{-1})$). The naive approach consists of choosing the two mutations for an individual according to the proposal distribution: if an individual has the genotype TT, we assign a mutation to each of their haplotypes, if it is tt, the wild type allele is assigned to each haplotype and in the case where tT is chosen a mutation is randomly assigned to one of the two haplotypes whiles the other haplotype will have the wild type allele. Then to calculate $Q(H_{-1}|H_0)$ we have by independence,

$$Q(H_{-1}|H_0)=\prod_{i=1}^{n} P(P_i|G_i).$$

Note that in the case of diploid data, it is not true that each all sequences are independent of one another as haplotypes are treated in pairs. Independence is assumed for the present, to simply calculations. This method has not yet been implemented but it is possible to do so in the future.

3.6.3 Adaptations of other researchers' methods for MapArg

Chapter 2 reviewed the fine mapping methods McPeek and Strah's (1999), Morris et al. (2002) and Zöllner and Pritchard (2005) with an emphasis on their treatment of the

penetrance parameters within their methods. The DHS method of McPeek and Strahs accounts for phenocopies (*i.e.* $f_0 > 0$), that are a result of multiple origins of the TIM and where the data are haploids. Incomplete penetrance ($f_1 < 1$) is not taken into account as the likelihood of haplotypes from the controls in the sample is not calculated within the coalescent model.

Similarly, the shattered coalescent model by Morris et al. (2002) deals only with phenocopy and does not look at incomplete penetrance. Although, the authors go one step further and account for sporadic events as well as multiple origins of the disease. This is quite important as some phenocopy may be caused by multiple origins of the disease; however, there are many complex diseases in existence that result from sporadic cases, *i.e.* cases who have no mutation in the ancestry of the sample. These sporadic cases are explained by environmental factors e.g. an individual may contract lung cancer from smoking (an environmental factor), without carrying a TIM for lung cancer. Although DHS and the shattered coalescent model appear to deal well with phenocopies, it is of interest to us to model both incomplete penetrance and phenocopy at the same time. Perhaps given the fact that the incomplete penetrance does not effect the performance of MapArg too strongly, it may be worth considering concentrating on modeling phenocopy only, in which case adaptations of the above methods could be considered.

The fine mapping method of Zöllner and Pritchard (2005), TreeLD, accounts for both incomplete penetrance and phenocopy for haploid data. The authors also mention that they are currently adapting their model to work for diploid data also. Recall that TreeLD constructs the genealogies independently of the phenotype data. The likelihood of the phenotype data is then estimated conditional on the genealogies. It is at this stage that a grid of possible levels for f_1 and f_0 is introduced, and the function $F = (f_1, f_0)$ that maximizes the phenotype data is chosen as the penetrance levels for the given sample. If f_0 and f_1 are unknown, a strategy similar to what is done in Zöllner and Pritchard (2005) can be used: we can integrate over a set of values. If we compare the $\ln L(\hat{r}_T)$ for the Cystic Fibrosis data, for $l = 8$ for example, when $f_1 = 1$ compared to $f_1 = 0.8$, we see the $\ln L(\hat{r}_T)$ is two times higher when $f_1 = 0.8$, suggesting that this

model is more probable.

Recall that as well as constructing genealogies independently of the phenotype data, TreeLD also constructs the genealogies independently of the mutation status. In order to model $Q(H_\tau)$ (the distribution upon which genealogies are constructed) in MapArg, it is necessary to know the status of the haplotype as being mutant or non-mutant.

CONCLUSION

Incomplete penetrance and phenocopy are two important phenomena that occur among populations with complex diseases. As MapArg assumed complete penetrance and no phenocopy to date, it was of great importance to see the effects, if any, that these parameters would have on the efficiency of this fine mapping method in finding the TIM. We have shown by way of simulation that incomplete penetrance does not appear to effect the performance of MapArg, whereas phenocopies among the sample render the method quite inefficient, even for quite low levels of phenocopy. The need to account for these parameters, especially phenocopy within the MapArg framework has become evident.

Given that the levels of penetrance and phenocopy are known *a priori*, two methods were developed in order to correct for the penetrance parameters. The first method, a rather straightforward approach, proved ineffective in most situations but improved efficiency under a few circumstances. However this method provides a starting point for the development of a model that works for diploid populations. Until now, MapArg works on haploid data but it is of interest to extend the method to diploid data.

Incorporating method 2 showed some improvement in the performance of MapArg. The most marked improvement can be seen when the sample size is increased. Also, the second method seems to work extremely well on the Cystic Fibrosis data, data that is known to have phenocopies resulting from multiple mutations. This result is very encouraging as it shows this method can work well for "real" data as well as simulated data, where situations are sometimes more ideal than in reality.

Further discussion as to how other methods of accounting for the penetrance parameters might somehow be adapted to suit the MapArg framework, is given. It is clear that there

remains a lot of further research in this area and it seems worthwhile to concentrate more on modeling phenocopy than penetrance as it is this parameter that has the greatest effect on MapArg.

BIBLIOGRAPHY

Baum, LE (1972) « An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process ». *Inequalities,* 3:1-8.

Felsenstein, J. (1981) « Evolutionary trees from DNA sequences: a maximum likelihood approach ». *J. Mol. Evolution,* vol. 17(6) p. 368–376.

Griffiths, R.C., and Marjoram, P. (1996a) « Ancestral Inference from samples of DNA sequences wih recombination ». *J. Comput. Biol.* vol. 3 p.479-502.

Griffiths, R.C., and Marjoram, P. (1996b) « An ancestral recombination graph ». *Progress in Population Genetics and and Human Evolution* (Donnelly P. and Tavaré S., Eds), vol. 87, p.257-270. Springer-Verlag New York.

Hudson, R. (2002) « Generating samples under a Wright-Fisher neutral model of genetic variation ». *Bioinformatics* vol. 18 p. 337-338.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A. and Buchwald, M. (1989) « Identification of the cystic fibrosis gene ». *Genetic analysis* Science vol. 245 p. 1073-1080.

Kingman, J.F.C. (1982) « The coalsecent ». *Stochast. Process Appl.* vol. 13 p. 235-248.

Larribe, F., Lessard, S. and Schork, N.J. (2002) « Gene Mapping via the Ancestral Recombination Graph ». *Theoretical Population Biology* vol. 62 p. 215-229.

Larribe, F. (2003) « Cartographie génétique fine par le graphe de recombinaison ancestral ». Thèse de doctorat, Montréal, Université de Montréal, 314 pages.

Larribe, F. and Lessard, S. (2007) « A Composite-condition-likelihood Approach for Gene Mapping Based on Linkage Disequilibrium in Windows of Markers ». *Submitted.*

McPeek, M.S. and Strahs, A. (1999) « Assessment of linkage disequilibrium by the decay of haplotype sharing with Application to Fine-Scale Genetic Mapping ». *Am. J. Hum. Genetics* vol. 65 p. 858-875.

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2002) « Fine-scale mapping of disease loci by hidden Markov models ». *Am. J. Hum. Genetics* vol. 74 p. 945-953.

Nordborg, M. (2001) « Coalescent Theory ». *Handbook of Statistical Genetics* Chapter 7 (Balding, D., Bishop, M. and Cannings, C. Eds), p. 179-212, Wiley, Chichester.

Olsen J.M., Witte J.S. and Elston R.C. (1999) « Tutorial in biostatistics. Genetic mapping of complex traits ». *Stat. Med.* vol. 18 p. 2961–2981.

Stephens, M. (2001) « Inference under the coalescent ». Chapter 8 (Balding, D., Bishop, M. and Cannings, C. Eds), p. 213-238, Wiley, Chichester.

Zöllner, S. and Pritchard, J.K. (2005) « Coalescent based association mapping and fine mapping of complex trait loci ». *Genetics* vol. 169 p. 1071-1092.