Modelling dependencies in genetic-marker data and its application to haplotype analysis

Michael T. Schouten

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2006

Abstract

The objective of this thesis is to develop new methods to reconstruct haplotypes from phase-unknown genotypes. The need for new methodologies is motivated by the increasing availability of high-resolution marker data for many species. Such markers typically exhibit correlations, a phenomenon known as Linkage Disequilibrium (LD). It is believed that reconstructed haplotypes for markers in high LD can be valuable for a variety of application areas in population genetics, including reconstructing population history and identifying genetic disease variants.

Traditionally, haplotype reconstruction methods can be categorized according to whether they operate on a single pedigree or a collection of unrelated individuals. The thesis begins with a critical assessment of the limitations of existing methods, and then presents a unified statistical framework that can accommodate pedigree data, unrelated individuals and tightly linked markers. The framework makes use of graphical models, where inference entails representing the relevant joint probability distribution as a graph and then using associated algorithms to facilitate computation. The graphical model formalism provides invaluable tools to facilitate model specification, visualization, and inference.

Once the unified framework is developed, a broad range of simulation studies are conducted using previously published haplotype data. Important contributions include demonstrating the different ways in which the haplotype frequency distribution can impact the accuracy of both the phase assignments and haplotype frequency estimates; evaluating the effectiveness of using family data to improve accuracy for different frequency profiles; and, assessing the dangers of treating related individuals as unrelated in an association study.

Acknowledgements

First and foremost, I would like to thank my two supervisors, Chris Williams and Chris Haley, for giving me every opportunity to thrive here. I would also like to thank Dr. Sarah Blott and Graham Plastow for valuable input, as well as Sygen International plc and the Biotechnology and Biological Sciences Research Council for funding.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Michael T. Schouten*)

Table of Contents

Abstract **i**

List of Figures **vi**

List of Tables **ix**

1 Introduction **1**

1.1 Genetics Background 1

1.2 The Importance of Haplotypes in Genetic Analysis 3

1.2.1 Traditional Importance in Linkage Analysis 3

1.2.2 Current Importance in Linkage-Disequilibrium Analysis 6

1.3 Motivation for Thesis 6

1.4 A Note on Human vs. Animal Genetics 7

1.5 Outline of the Thesis 8

List of Symbols, Acronyms and Notation **1**

2 Haplotype Reconstruction for Unrelated Individuals **10**

2.1 Overview 10

2.2 Comparison of Modelling Assumptions 12

2.2.1 The EM algorithm for Unrelated Individuals (HW-EM) 13

2.2.2 The Naive Gibbs Sampler (HW-GS) 13

2.2.3 The Coalescence-Based Gibbs Sampler (C-GS) 14

2.3 Computational Challenges 15
2.3.1 Accommodating Large Numbers of Markers 16
2.3.2 Convergence Rates of the Gibbs Samplers 18
2.4 Model Performance 18
2.5 Looking ahead: Extensibility of the EM Algorithm 20

3 Haplotype Reconstruction for Pedigrees 21
3.1 Overview 21
3.2 Stochastic Methods for Haplotype Reconstruction on Pedigrees 22
3.2.1 The Lander-Green Algorithm 22
3.2.2 Applications and Robustness of the Lander-Green Algorithm 23
3.3 Deterministic Methods for Haplotype Reconstruction on Pedigrees 26
3.4 Simulation Study: The Need for a New Paradigm 26

4 A Unified Model for Haplotype Reconstruction 29
4.1 Overview 29
4.1.1 Notation and Assumptions 32
4.2 The Graphical Model Paradigm 33
4.2.1 Model Specification and Visualization 34
4.2.2 Inference 35
4.3 An EM Algorithm for Outbred Half-Sib Pedigrees 39
4.3.1 Specifying the Complete-Data Log-Likelihood 40
4.3.2 Calculating the Marginal Distributions of the Latent Data 40
4.3.3 Evaluating the Complexity of the Algorithm 42
4.3.4 Extensions to the Graphical Model 43

5 Simulation Studies 46
5.1 Overview 46
5.2 Simulation Study: The Effectiveness of Pedigree Data in Haplotype Reconstruction 47
5.2.1 Specifying the Haplotype Frequency Distributions 48

5.2.2 Simulating the Data 49
5.2.3 Summarizing the Results 50
5.2.4 Impact of Family Data on Haplotype Frequency Estimation 51
5.2.5 Impact of Family Data on Phase Reconstruction Accuracy 53
5.3 Simulation Study: Sampling from Small Hierarchical Populations 56
5.3.1 Overview 56
5.3.2 The Sampled Population 57
5.3.3 Simulation Strategy 59
5.3.4 Results 60
5.4 Discussion 61

6 Haplotype Analysis in Association Studies **64**
6.1 An Empirical Assessment of Association Studies 65
6.1.1 Failure to Detect Disease Variants 65
6.1.2 The High False-Positive Rate 69
6.2 Using Haplotype Data in a Mapping Analysis 71

7 Conclusions and Future Work **74**
7.1 Summary 74
7.2 Future Work 76

A Optimal Experimental Design Revisited **78**

B Multimodality **82**
B.1 Deriving the Likelihood 83
B.2 Results 84

C Estimating Haplotype Frequencies Using the EM Algorithm **86**

Bibliography **94**

List of Figures

1.1 Simple characterization of marker sequencing technology. 3

1.2 The relationship between phase and genotype data. 4

3.1 Lander-Green Representation of the Likelihood for a Parent-Child Trio at two loci. 24

4.1 DAG for the joint distributions relevant to haplotype reconstruction for unrelated individuals (left) and half-sib pedigrees (right). 36

4.2 Deriving the Junction Tree for a Half-Sib Pedigree, where $\eta_i = 2$ From top left: (1) the original DAG; (2) the moralized (undirected) graph and (3) the Junction Tree. The cliques, which are read off the moralized graph, are defined as follows: $C_1 = \{s_i, d_{1i}, z_{1i}\}, C_2 = \{s_i, d_{2i}, z_{2i}\}, E_1 = \{z_{1i}, y_{1i}\}$ and $E_2 = \{z_{2i}, y_{2i}\}$. 37

4.3 The Junction Tree for a half-sib pedigree of size η_i. The cliques are defined as: $C_k = \{s_i, d_{ki}, z_{ki}\}$, $E_k = \{z_{ki}, y_{ki}\}$. 39

4.4 DAG for the joint distributions relevant to haplotype reconstruction for unrelated individuals (left) and half-sib pedigrees (right) when parental genotypes are included. 44

4.5 Deriving the Junction Tree for a Full-Sib Pedigree. From top left: The original DAG; the moralized (undirected) graph; and, the Junction Tree. . . 45

5.1 Conceptual Framework for Evaluating Population Structure and Relevant Population Parameters of a Standard Breeding Design 57

5.2 Impact of Sample Size, Number of Sires and Model Specification on Discrepancy of Θ. 62

A.1 Gaussian Graphical "Sibship" 79

B.1 Plot of log-likelihood against α for two-locus genotype counts (n) for *Idh1* and *Mdh* loci in mosquito data presented in Weir (1990). 85

B.2 Plot of log-likelihood against α for different counts of double heterozygotes (n_5). All other genotype counts are held constant at 10. 85

List of Tables

2.1 Summary of phase reconstruction error rates from various methods. 19

3.1 Simple example illustrating dangers of assuming LE for markers in tight LD when reconstructing haplotypes 25

3.2 Comparison of Haplotype Reconstruction Accuracy for Sires in Half-Sib Pedigrees using Three Different Models. 28

5.1 Summary statistics for haplotype frequency estimates used in data analysis. 49

5.2 Categories of family data that can be included with a sample of phase unknown genotypes in simulation study. 50

5.3 Impact of family size, family information and parental haplotype reconstruction accuracy 52

5.4 Percent reduction in discrepancy for Θ^0 (left) and Θ (right) attributable to modelling dependencies in random population sample. 61

6.1 Impact of treating half-sibs as unrelated on two nonparametric tests. 70

B.1 Components of the log-likelihood for a sample of phase-unknown genotypes for two biallelic loci. 83

C.1 Haplotypes for two biallelic loci. 86

C.2 A Sample of Three Phase-Known Genotypes Used In Example 1. 87

C.3 Maximum Likelihood Estimates of Θ for Data Set Given in Example 1. . . 88

C.4 A Sample of Three Genotypes which are Completely Informative For Phase. 88

C.5 Initial Haplotype Frequency Estimates (Θ^0) 89
C.6 Updated Haplotype Frequency Vector after One Iteration of the EM Algorithm (Θ^1). 89
C.7 The Sample of Three Genotypes Used in Example 3. 90
C.8 Updated Haplotype Frequency Vector after One Iteration of the EM Algorithm (Θ^1). 91
C.9 Updated Haplotype Frequency Vector after the Second Iteration of the EM Algorithm (Θ^2). 92

Chapter 1

Introduction

The objective of this thesis is to develop new statistical methods for haplotype reconstruction. The first part of this introduction provides the necessary genetics background and theory to make the thesis self-contained. This includes defining what haplotypes are and why they need to be reconstructed *in silico*. The next section describes their importance in genetic analysis and speculates on why there might be the need for new reconstruction algorithms. The final section provides an outline for the remainder of the thesis.

1.1 Genetics Background

This section provides the necessary genetics background and terminology to make the thesis self contained. It begins by characterizing the genome in way that will be meaningful to information scientists, and then providing the relevant terminology. All statistics pertain to the human genome.

An appealing feature of genetic marker analysis is that many relevant biological concepts can be readily understood by the information theorist. This is because the genome, which is the collection of all heritable information, is naturally characterized as a pair of strings (one inherited from each parent). Each string is 3.6 billion characters long over the alphabet A,C,T and G. The information encoded on the two strings contain both the necessary and sufficient information to predict all completely heritable characteristics for an individual.

Each character at given location (or locus) is called an allele. A pair of alleles at a given locus is a genotype. If the genotype is comprised of the same alleles, it is homozygous. If it contains different alleles, it is heterozygous.

Any observed, physiological characteristic that has a genetic basis is called the phenotype. Hence, phenotypic variation across individuals that cannot be attributed to the environment can be attributed to genotype variation in the genome.

It is now established that 99.9% of the genome is monomorphic, i.e. all individuals in the population will be homozygous for the same alleles. Hence, genetic variation between two individuals can be attributed to differences in less than 1% of the genome. Loci that are not monomorphic are polymorphic, also referred to as (genetic) markers. The markers that will be considered in this thesis are Single Nucleotide Polymorphisms (SNPs). SNP markers typically feature exactly two alleles that are segregating in a population, and for the remainder of this thesis SNPs will be assumed biallelic. Hence, it is common to label alleles as 0 or 1.

The genome has been characterized as a pair of very long strings of alleles. The information is actually divided across 23 substrings, or chromosomes. Any sequence of alleles on a given chromosome (string) is called a haplotype. Just as a pair of alleles is referred to as a genotype, a pair of haplotypes is referred to as the phase. Much has been made of the sequencing technology that allowed the human genome to be mapped. When sequencing an individual for a set of markers, it is natural to envision that the phase is returned, as depicted in the top of Figure 1.1. This type of technology is very expensive and is only available in very limited contexts. Instead, the standard genetic information that is provided for marker data are the genotypes for each locus.

The resulting problem, as shown in Figure 1.2, is that while phase information is completely informative for genotypes, the converse is not true. Specifically, the number of phase configurations that can resolve a set of phase unknown genotypes increases exponentially with the number of heterozygous loci. The objective of haplotype reconstruction is to use all available information to determine which is the correct one.

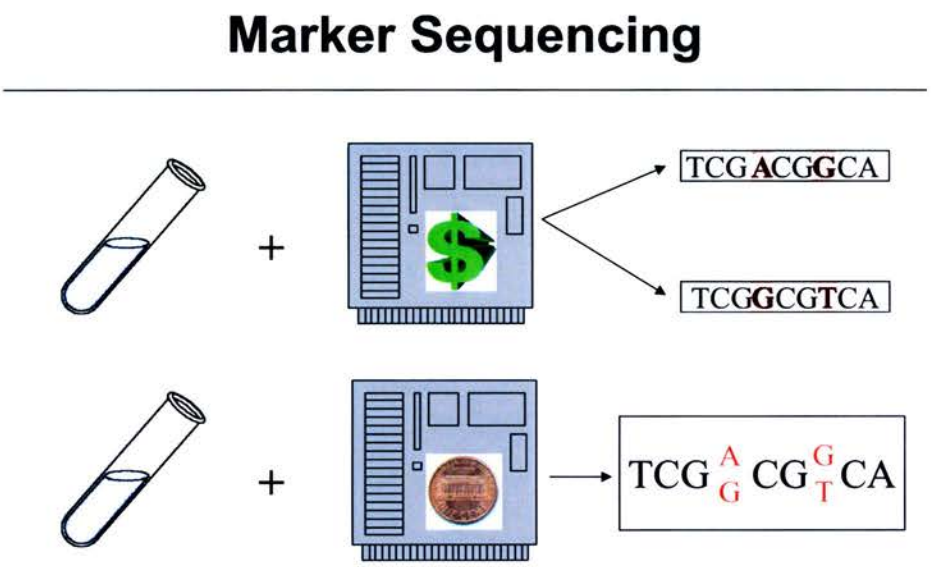


Figure 1.1: Simple characterization of marker sequencing technology. Individual's DNA is extracted and submitted to a black-box sequencing machine. (Top) Phase is returned. (Bottom) Genotypes are returned, where single characters denote homozygotes.

1.2 The Importance of Haplotypes in Genetic Analysis

Haplotypes play a central role in many types of genetic analysis. Traditionally, reconstructed haplotypes were essential in linkage mapping, which was basis of the Human Genome Project. Currently, the interest in haplotypes stems from the availability of markers in high linkage disequilibrium. The relevance of haplotypes in each of these contexts is now discussed.

1.2.1 Traditional Importance in Linkage Analysis

The basis of linkage analysis lies in how parents transmit genetic information to their progeny. Each individual has two sets of chromosomes, one inherited from each parent.

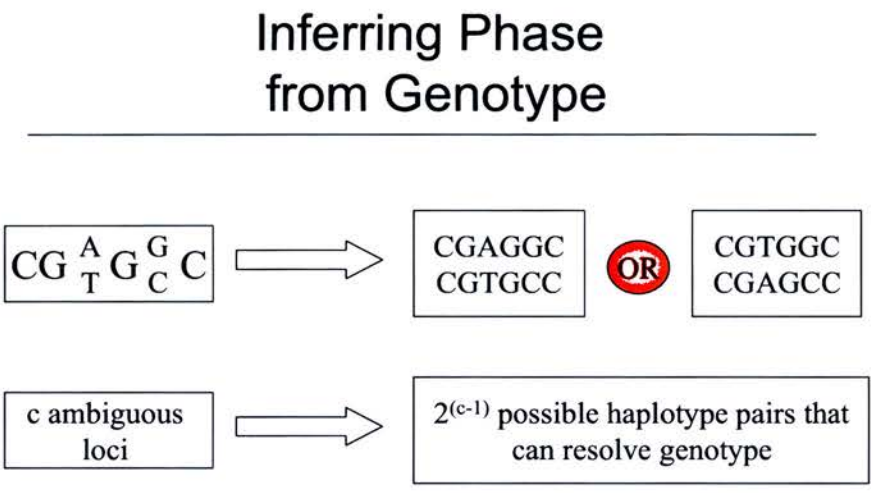


Figure 1.2: The relationship between phase and genotype data.

Equivalently, the progeny inherits one chromosome from a given parent. The chromosome that is passed from parent to progeny is called a gamete, and features a blend of alleles from each of the parent's two chromosomes (i.e. the gamete will feature a blend of the progeny's grandparents' alleles).

The process of gamete formation is called meiosis. One of Mendel's seminal contributions was his attempt to explain the process in terms of two simple laws. The first law, the law of independent segregation, stipulates that each allele has an equal chance of being transmitted from parent to progeny (i.e. that the progeny has equal chance of inheriting either grandparent's allele). The second law, the law of independent assortment, states that alleles at different loci are transmitted independently. While Mendel's first law is correct, the second is only true if the two alleles reside on different chromosomes. Alleles on the same chromosome are transmitted more frequently than chance would allow, and are therefore defined as linked.

The reason why Mendel's second law does not hold is that during gamete formation, pairs of homologous chromosomes line up and several zones of contact (crossover points) are established. The number of crossovers is very small (in the human there is only an average of three per chromosome). The alleles, or haplotype, between two crossover points are transmitted as a unit, while haplotypes in successive intervals are transmitted independently. It is clear, therefore, that the closer two markers are on the same chromosome, i.e. the more tightly linked two alleles are, the more likely that constituent alleles will be transmitted as a unit.

Haplotypes that exhibit a mixture of grandmaternal and grandpaternal alleles are defined as recombinant. Equivalently, recombination has occurred if a progeny's haplotype consist of alleles that are not identical by descent (IBD) with the same grandparent.

The identification of recombinant and nonrecombinant haplotypes is the cornerstone of genetic mapping. When a marker is first discovered, its position in the genome is unknown. Linkage analysis refers to a group of statistical methods to evaluate whether the number of nonrecombinant haplotypes featuring this newly discovered marker is significantly larger than the number of recombinant haplotypes. If the difference is significant, linkage between this marker and other markers on the haplotype is indicated.

Creating a linkage map of genetic markers (i.e. polymorphic loci) is an essential step for meiotic mapping of genes. Meiotic mapping methods attempt to map the location of a gene by inferring how a genetic marker segregates with a phenotype. Creating a linkage map was thus one of the primary objectives of the human genome project. However, a linkage map will only provide the relative distance between two polymorphic markers. In a linkage map, the distances between two markers centimorgans (cM), where 1 cM equals a 1% chance that a marker at one genetic locus will be separated from a marker at another locus due to crossing over in a single generation. Physical maps, by contrast, show the exact location of genetic markers, and the distance between them measured in base pairs. Each map is important in validating the other, and mapping the genome for a species will entail creating both kinds of maps.

1.2.2 Current Importance in Linkage-Disequilibrium Analysis

Alleles that are in linkage disequilibrium are not independent within a population. Specifically, consider two biallelic markers, the first with allele frequencies p_1 and p_2, and the second with allele frequencies q_1 and q_2. Let P_{11} be the frequency of the haplotype formed by the first allele of each marker. LD occurs when

$$P_{11} - p_1 q_1 \neq 0.$$

On average, markers in close physical proximity are expected to exhibit higher LD than those spaced farther apart[1]. The reason is that any given variant must have been introduced into the population by some individual (a founder). That variant was introduced on a founder haplotype, and each successive generation, the size of that haplotype diminished as the number meiotic events increased. This is why older populations exhibit greater haplotype diversity. Haplotype analysis has therefore provided invaluable support to establishing that Africa is the oldest population (the "Out of Africa" hypothesis) by demonstrating that African populations exhibit greater haplotype diversity than other major world populations(Tishkoff et al., 1996).

LD-based analysis is becoming increasingly popular given the advent of fine-scale genotyping technology. Haplotypes of tightly linked markers have already proven valuable in reconstructing evolutionary history of several species. There is widespread hope that haplotypes may be equally valuable in other contexts, notably complex disease mapping.

1.3 Motivation for Thesis

As noted above, there is currently a strong interest in how best to use LD information for fine-scale mapping and association analysis of complex traits. A growing number of studies demonstrate that haplotype-based approaches may provide more power and accuracy in locating causative disease variants than single-locus methods (see, e.g. Zhao et al., 2003;

[1]This has been supported by empirical evidence, despite the fact that LD can be induced by a myriad of other factors that can obscure this relationship.

Morris et al., 2002; Fallin et al., 2001). These haplotype-based studies commonly follow a two-step procedure: first, haplotypes are inferred from a sample of phase-unknown genotypes using a computational algorithm, and second, inferred haplotypes are fed into a multi-locus LD model, where they are treated as having been directly observed.

There are two approaches to inferring haplotypes from population data, both with potential drawbacks. One approach is to use family data, which may be able to deterministically resolve phase for genotypes featuring multiple heterozygous loci. However, ascertaining this information can be costly. Furthermore, many popular pedigree-based methods assume Linkage Equilibrium, which is cause for concern.

A second approach is to infer haplotypes directly from population data. A variety of statistical algorithms exist for random-mating populations, and good comparative surveys are available (see, e.g. Stephens and Donnelly, 2003; Zhao et al., 2003). A problem with reconstructing haplotypes using these models is that there may be considerable uncertainty associated with the inferred haplotypes.

The objective of the thesis is to develop a robust statistical model that can accommodate two kinds of dependencies: dependencies across markers (LD) and dependencies across individuals (family data) and to assess the importance of each in the quality of haplotype reconstruction.

1.4 A Note on Human vs. Animal Genetics

This thesis was developed in the context of animal breeding. This is why the half-sibship is the pedigree structure that is the basis for the model in Chapter 4. It is also why males and females are referred to as sires and dams respectively. However, most of the literature that this thesis utilizes was published in the context of human genetics, which is where the vast amount of theoretical and empirical research in the areas of haplotype reconstruction and LD based analysis has been conducted.

It is not inappropriate to apply results from human genetics to outbred animal stock. The thesis demonstrates the importance of accounting for LD in haplotype reconstruction. LD

profiles for livestock populations are only now emerging (see, e.g. Heifetz et al., 2005). While LD is likely to vary widely across the genome in livestock populations (as it does with humans), it appears that, on average, LD will be more extensive in livestock populations than in humans.

1.5 Outline of the Thesis

- **Chapter 2** examines haplotype reconstruction algorithms for unrelated individuals. In particular, this chapter evaluates three of the most popular algorithms and concludes that the EM-based approach is a sound method in terms of accuracy, accessibility and extensibility.

- **Chapter 3** reviews haplotype reconstruction algorithms for pedigrees and concludes that these are not relevant for analyses featuring tightly linked markers.

- **Chapter 4** introduces a model to reconstruct haplotypes for unrelated individuals, family-child trios and arbitrarily large half-sib pedigrees. This model is efficient over non-recombinant regions and, crucially, accommodates LD between loci.

- **Chapter 5** is devoted to simulating and analyzing results. Simulations are conducted for a diverse set of haplotype frequency distributions, all of which have been previously published in empirical studies. A wide variety of important results regarding the effectiveness of using pedigree data in a population study are presented in a coherent, unified framework. Insight is provided into the different properties of the haplotype frequency distribution that can influence experimental design. It is shown that a preliminary estimate of the haplotype frequency distribution can be valuable in large population studies with fixed resources.

- **Chapter 6** provides a critical evaluation of the role of haplotypes for fine-mapping studies, including simulation studies to illustrate potential problems using haplotypes for ANOVA (model-free) mapping studies.

- **Chapter 7** features concluding remarks and proposes some directions for future research.

Chapter 2

Haplotype Reconstruction for Unrelated Individuals

‘ Haplotype reconstruction algorithms for unrelated individuals are primarily likelihood based, where the likelihood reflects Hardy-Weinberg assumptions. This chapter evaluates three of the most popular algorithms and concludes that a popular EM-based approach is a sound method in terms of accuracy, accessibility and extensibility. These conclusions inspired the model in Chapter 4, which can be regarded as a generalization of this approach to pedigree data.

2.1 Overview

This chapter provides a comparative analysis of three of the most popular likelihood-based approaches to haplotype reconstruction for unrelated individuals. The first is an EM-based approach that has been developed and evaluated in many different contexts (see, e.g. Hill, 1974; Terwilliger and Ott, 1994; Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Fallin and Shorck, 2000; Kirk and Cardon, 2002; Qin et al., 2002). The second is a Bayesian model introduced by Niu et al. (2002). The method specifies a Dirichlet prior over the haplotype frequency distribution, and this is then combined with the same complete-data likelihood evaluated by the EM algorithm. Although this method is commonly referred

to as HAPLOTYPER, it has also been referred to as the "Naive" Gibbs Sampler since the Dirichlet prior results in sampling behavior that is inconsistent with basic population genetics theory (Stephens and Donnelly, 2003). This description is preferable since it helps underscore the key differences between these methods. The final method, which is also Bayesian, was introduced by Stephens et al. (2001) and employs an approximate coalescence prior which suggests an improvement over the method by Niu (2004), provided that the coalescent is a reasonable model for the data.

Despite the ubiquity of these models, there remains a considerable amount of confusion over relative benefits/drawbacks for using each one. As suggested by Stephens and Donnelly (2003), a critical first step to understanding the these (or any other) statistical models is to distinguish between the modelling assumptions and the computational complexity. Section 2.2 provides a mathematical and qualitative description of these modelling assumptions, beginning with the derivation of the likelihood of the observed data that is common to all three. Briefly, the key assumption for both the EM-based approach and the Gibbs Sampler proposed by Niu et al. (2002) is that the population is in HWE. Throughout the chapter, we therefore refer to these models as HW-EM and HW-GS, which allows us to distinguish between the probability model and the model of inference. Similarly, we refer to the method of Stephens et al. (2001) as C-GS to distinguish between the probability model (coalescence-based evolution in addition to HWE) and the method of inference (Gibbs Sampling).

Section 2.3 discusses several important computational problems facing all three methods. Among the more important conclusions is that the complexity of all three algorithms is the same, and a popular heuristic that deals with the resulting limitations on the number of markers can be adapted to each of the methods. Section 2.4 discusses the relative performance of each the three algorithms. This provides evidence that, for many marker configurations that are used in fine-scale genetic analysis, the HW-EM algorithm performs as well as the two MCMC methods in small samples.

2.2 Comparison of Modelling Assumptions

As noted above, all three methods, are based on the same likelihood function, which reflects the assumption that the population is in Hardy-Weinberg Equilibrium (HWE). While describing this likelihood, we also introduce some key notation that will be used throughout the thesis.

We are considering a candidate region in the genome characterized by L tightly linked biallelic loci. Let $\mathbf{h} = h_1 \dots h_M$ denote the $M = 2^L$ possible haplotypes, and let $\Theta = (\theta_1, \dots, \theta_M)$ denote corresponding haplotype frequencies in the target population[1].

For a large, panmictic population, we can specify the probability of observing a given phase configuration, $z = (h_i, h_j)$ as

$$p(z = h_i, h_j | \Theta) = c_{ij}\theta_i\theta_j \tag{2.1}$$

where

$$c_{ij} = \begin{cases} 2, & i \neq j \\ 1, & \text{otherwise.} \end{cases}$$

Similarly, the probability of observing a given phase unknown genotype, y, in a panmictic population is:

$$p(y|\Theta) = \sum_{z \in \mathbf{z}(y)} p(z|\Theta), \tag{2.2}$$

where $\mathbf{z}(y)$ is the set of all possible phase configurations that can resolve y.

Let $\mathbf{y} = y_1 \dots y_N$ denote a sample of N phase unknown genotypes. For a given haplotype frequency profile, Θ, the log-likelihood of the observed data is thus:

$$\log p(\mathbf{y}|\Theta) = \sum_{i=1}^{N} \log \left[\sum_{z \in \mathbf{z}(y_i)} p(z|\Theta) \right] + \text{Constant.} \tag{2.3}$$

In general, the summation in between the brackets of equation (2.3) will prevent an analytic solution for the maximum likelihood estimate of Θ. However, the problem is amenable to estimation using any data augmentation approach.

[1] More precisely, θ_i represents the proportion of h_i haplotypes that segregate in the population.

2.2.1 The EM algorithm for Unrelated Individuals (HW-EM)

The EM algorithm is the most straightforward data augmentation method. The general approach entails augmenting each y_i, i.e. the observed phase-unknown genotype for each member in the sample, by the corresponding phase configurations. We denote these latent phase configurations by $\mathbf{z} = z_1 \ldots z_N$. Note that since $\mathbf{z}$ is completely informative for $\mathbf{y}$, the augmented likelihood, $p(\mathbf{y},\mathbf{z}|\Theta)$ is equivalent to $p(\mathbf{z}|\Theta)$, which follows a multinomial distribution. The expected log-likelihood of the augmented data which can therefore be expressed as:

$$E_{p(\mathbf{z}|\mathbf{y}\tilde{\Theta})} \log[p(\mathbf{y},\mathbf{z}|\Theta)] \;=\; \sum_{i=1}^{N}\sum_{j=1}^{M} E_{p(z_i|y_i,\tilde{\Theta})} n_{ij} \log \theta_j + \text{Constant}, \tag{2.4}$$

where $\tilde{\Theta}$ denotes the current estimate of Θ and n_{ij} refers to the number of times haplotype j appears in the phase configuration of individual i. Once (2.4) has been calculated, the result is maximized with respect to Θ and the process is repeated until $\tilde{\Theta}$ converges at a maximum, $\hat{\Theta}$. To illustrate the computations involved in this process, a simple example has been provided in Appendix 3.

It is important to note that while the statistical objective of the EM-based approach to haplotype reconstruction is to calculate the maximum likelihood estimate of the haplotype frequencies ($\hat{\Theta}$) given genotype data ($\mathbf{y}$), these frequency estimates can then be used to reconstruct phase probabilities using $p(z|y,\hat{\Theta})$. Hence the HW-EM algorithm is considered appropriate for experiments requiring both haplotype frequency estimates and phase calls.

2.2.2 The Naive Gibbs Sampler (HW-GS)

As stated above, the likelihood for the augmented data, $p(\mathbf{y},\mathbf{z}|\Theta)$, follows a multinomial distribution. The method of Niu et al. (2002) begins by introducing a Dirichlet (conjugate) prior, which is conjugate to the multinominal distribution, for Θ, i.e. $\Theta \sim \text{Dirichlet}(\beta)$ where $\beta = (\beta_1, \ldots, \beta_M)$. Unlike the HW-EM algorithm, HW-GS does not estimate Θ directly. Instead, $p(\mathbf{y},\mathbf{z})$ is then derived by integrating out Θ from the full joint distribution,

from which phase call estimates ($\mathbf{z}$) are obtained by Gibbs Sampling. Haplotype frequency estimates can be obtained by averaging the MAP estimates of $\mathbf{z}$.
Niu et al. (2002) show that the relevant conditional distributions can be expressed as:

$$p(z_i = h_j, h_k | \mathbf{z}_{-i}, \mathbf{y}) \propto (n_j + \beta_j)(n_k + \beta_k), \tag{2.5}$$

where $\mathbf{z}_{-i}$ represents the haplotype pairs for all subjects excluding the ith individual, and n_j and n_k denote the counts of haplotypes j and k that are in $\mathbf{z}_{-i}$.
Niu et al. (2002) suggest an prior-annealing strategy that involves choosing large values of β at the beginning of each iteration and then gradually decreasing them as the iteration progress. Their motivation is to allow the Gibbs sampler to freely manoeuvre in haplotype space without getting stuck on a local maximum.
It must be stressed that the values for β are chosen solely on the basis of computational considerations. They do not incorporate any population-specific (prior) knowledge regarding the the haplotype distribution. As a result, this sampling procedure often exhibits behavior that is inconsistent with basic population genetics theory. This "naive" behavior is best understood when considering a uniform prior on the haplotype frequency distribution is used (i.e. $\beta = \mathbf{1}$). For small sample size, none of the haplotypes in the candidate phase configurations may appear in the rest of the sample ($\mathbf{z}_{-i}$), i.e $n_j = 0$ and $n_k = 0$ for all relevant haplotypes. The HW-GS would then assign equal weight to all candidate phase configurations. The standard neutral (or coalescent) theory of evolution stipulates that haplotypes are more likely to be clustered, i.e. we would want to place more weight to those haplotypes which look similar to ones that have already been assigned in $\mathbf{z}_{-i}$. It is unclear how to achieve this in a principled way with a Dirichlet prior.

2.2.3 The Coalescence-Based Gibbs Sampler (C-GS)

As discussed in the previous section, an informed prior would place greater weight on those haplotype configurations that look similar to those that we are conditioning on. Stephens and Donnelly (2000) derive just such a sampling distribution, where similarity is based on the coalescent theory. The model predicts the configuration of a single haplotype (h)

conditioned on a set of previously sampled haplotypes,(H), according to the following distribution:

$$p(h|H) \propto \sum_{\alpha} \sum_{s=0}^{\infty} \frac{r_\alpha}{N} \left(\frac{\mu}{\mu+N} \right)^s \frac{N}{\mu+N} (P^s)_{\alpha h}, \tag{2.6}$$

where r_α is the number of haplotypes of type α in H, N is the total number of sampled haplotypes, μ is a mutation rate and $(P^s)_{\alpha h}$ is the probability that a haplotype of type α could result in h after mutational events described by s. Stephens et al. (2001) state that "informally, this corresponds to the next sampled haplotype, h, being obtained by applying a random number of mutations s to a randomly chosen existing haplotype, α."
This distribution can easily be applied to haplotype reconstruction since equation (2.5) can be rewritten as

$$p(z_i = h_j, h_k | \mathbf{z}_{-i}, \mathbf{y}) \propto p(h_j | \mathbf{z}_{-i}, \mathbf{y}) p(h_k | \mathbf{z}_{-i}, h_j, \mathbf{y}) \tag{2.7}$$

These conditional distributions are inconsistent since they were not derived from a proper joint distribution, and this model has been referred to as a "pseudo" Gibbs Sampler (perhaps in retaliation for labeling the previous model as "naive"). As discussed in Stephens and Donnelly (2003), inconsistent distributions can be problematic because Gibbs sampling is not guaranteed to converge to the joint distribution. However, Gibbs Sampling on these distributions is guaranteed to converge since the Markov Chain is irreducible and aperiodic. Stephens and Donnelly (2000) also show that as $N \to \infty$, the sampling distribution converges to one that is consistent with Hardy-Weinberg assumptions. Equivalently, for large samples, the HW-GS and C-GS exhibit the same performance. The problem of quantifying what constitutes a large sample is addressed in section 2.4.

2.3 Computational Challenges

It is well known that the number of loci (rather than the sample size) can impede computational tractability. This section demonstrates that computational complexity with respect to the number of loci of the MCMC methods is equivalent to the EM algorithm. One problem that is specific to MCMC methods is the difficulty in determining convergence criterion.

This section also discusses the practical implications of this problem in the context of haplotype reconstruction.

2.3.1 Accommodating Large Numbers of Markers

It is well known that the HW-EM algorithm can only accommodate a limited number of loci (≤ 20). This can attributed to the requirement of enumerating all phase configurations that are consistent with a given genotype (i.e. calculating $p(z_i|y_i, \tilde{\Theta})$ during the E-Step). Both models that use MCMC methods (specifically, Gibbs Sampling) were introduced as being able to accommodate arbitrarily large numbers of loci, and a resulting misconception is that the MCMC methodology itself improves on limitations of the HW-EM algorithm. The computational complexity of all models are, however, equivalent since both methods that utilize Gibbs Sampling also require the enumeration of all possible phase configurations that are consistent with a given phase-unknown genotype (see equations 2.5 and 2.7 for the HW-GS and C-GS respectively).

The complexity of computing $p(z_i|y_i, \tilde{\Theta})$ increases linearly with the number of possible phase configurations that can resolve an ambiguous genotype. However, the number of possible phase configurations increases exponentially with the number of heterozygous genotypes, which limits the number of SNPs that can reasonably be evaluated. This constraint will apply to *any* Gibbs Sampling approach that requires estimating the distribution of phase given marker data.

Niu et al. (2002) describe a heuristic to accommodate large numbers of loci, which they call partition-ligation (PL). PL is completely generic in that in can be applied to any algorithm that infers haplotype frequencies. It has since been incorporated into the C-GS (Stephens and Donnelly, 2003) and the HW-EM algorithm (Qin et al., 2002). It has also been proposed for related pedigree-based haplotype reconstruction methods (Abecasis and Wigginton, 2005; Schouten et al., 2005).

The heuristic employs a divide-and-conquer strategy. First, the L loci are partitioned into smaller tractable "atomistic units" (i.e. subsets where all constituent haplotypes can be exhaustively enumerated). The haplotype reconstruction algorithm is then applied to each

unit, but only the B most probable haplotypes are recorded. Pairs of atomistic units are merged (ligated), by concatenating all B^2 possible haplotype combinations.

It is clear that a sufficiently high value of B should be chosen so that correct haplotypes are not discarded. Setting B to 40 is regarded as sufficiently large for atomistic units less then 7, however no formal analysis has been conducted to substantiate this. Until further research is conducted, PL should be treated in the same way as multimodality, i.e. the same data should be analyzed under different parameter settings to ensure a global optimum is reached.

It should be noted that PL-EM is not the only method that has been introduced to cope with a large number of loci. Thomas (2003) also proposes a recursive algorithm, where, at each stage, the list of markers is split in half so that the base case consists of analyzing two locus haplotypes. The novel insight is to run the HW-EM algorithm for one iteration on each set of markers and eliminate those haplotypes that have an estimated frequency of zero. The assumption is that it is not necessary to wait until the HW-EM algorithm converges before many of the haplotypes with zero frequency will have been estimated. A key distinction between this approach and PL-EM is that this method will not discard any haplotypes with positive frequency (recall that PL-EM will only choose the B most likely haplotypes at any stage of the recursion). Hence, when many haplotypes are segregating in the population (or when the level of missing data is high), the algorithm may be prohibitive.

Clayton (2002) introduced an even simpler approach to culling haplotypes with zero frequency. The software, SNPHAP, starts by fitting two-locus haplotypes and extends the solution one locus at a time. Clayton (2002) concedes that the order in which loci are introduced can result in different solutions and recommends running SNPHAP using different marker orders, as well as culling haplotypes after every k loci (k is not specified) rather than after every single locus.

In conclusion, more work needs to be done to identify the relative merits of these approaches. PL-EM is becoming increasingly popular as it has been applied to other haplotype reconstruction algorithms, such as Stephens and Donnelly (2003) and Zhang et al. (2005).

2.3.2 Convergence Rates of the Gibbs Samplers

An important problem in Gibbs sampling is determining the appropriate burn-in period, i.e. the number of initial iterations that must be discarded before the Markov Chain has converged to the desired stationary distribution. The convergence time for the coalescence-based model is much longer than for the naive approach. Specifically, while 5,000 updates are considered acceptable for the HW-GS to converge to the posterior distribution, 2,000,000 updates are needed to provide a reasonable approximation to the posterior defined by the C-GS (Stephens and Donnelly, 2003). The practical implications of this may be profound: A recent study by Niu (2004) alleges that it would take months to complete a standard simulation study for 500 subjects using C-GS, while the HW-GS could easily accommodate this level of data.

2.4 Model Performance

In this section, we examine the results of several comparative studies that evaluate the accuracy of the three methods described above. These studies are primarily concerned with comparing phase reconstruction accuracy. However, as we examine in Chapter 5, phase is determined as function of the haplotype frequency distribution. Hence, a model can be evaluated in terms of the haplotype frequency distribution.

When considering relative performance of these three models, it is important to remember that the likelihood of all three models is equivalent. Hence for a "large" (or sufficiently informative) sample sizes the performance should be equivalent. A reasonable, question, therefore, is: when is a sample sufficiently informative to render the prior distribution irrelevant? The information content of a sample if influenced by both the heterozygosity[2] and sample size. While it is impossible to test all combinations of these parameters, some meaningful results are available. Fallin and Shorck (2000) demonstrate that for a five locus system, samples sizes that are ≥ 50 are sufficient to guarantee a near-zero MSE for all

[2]The hererozygosity is defined as $1 - \sum_i \theta_i^2$ and is the expected number of heterozygotes in the sample.

Gene (Loci,Haplotypes,Sample Size)	EM/HW-GS	C-GS
CFTR (23,56,28)	.40	.47
ACE (52,n/a,11)	.19	.18
β_2AR (23,10,15)	.09	.05
β_2AR (23,10,121)	0	0

Table 2.1: Summary of phase reconstruction error rate from comparative studies evaluating HW-EM, HW-GS and C-GS. EM and HW-GS are grouped together since their performance is consistently similar. Error rates for β_2AR (23,10,121) from Niu et al. (2002). All other data from Stephens and Donnelly (2003).

haplotype frequency vectors[3]. Zhang et al. (2001) show that the performance of PHASE and HW-EM is equivalent for a wide variety of two locus systems featuring sample sizes of ≈ 30, which indicates that this sample size/marker combination is sufficient to achieve good estimates of Θ. These studies are quite useful since haplotype-based mapping analysis use haplotypes based on two or five SNPs (see, e.g. Grapes et al., 2004; Zaykin et al., 2001).

The comparative studies that examine all three algorithms typically feature extremely small sample sizes (Stephens et al., 2001; Niu et al., 2002; Qin et al., 2002; Stephens and Donnelly, 2003). The heterozygosity is not reported, but can be crudely approximated by the number of loci and the number of haplotypes that are segregating in the system. A summary of the phase reconstruction error rate for the data analyzed by the four studies cited above are given in Table 2.1. Based on the evidence published thus far, we conclude that the MCMC approaches do not offer any clear advantages over the HW-EM algorithm, and that we can feel confident using it for the kinds of data that are typically used in fine-mapping

[3]It is evident that parameter estimation improves with sample size. Equivalently, the MSE will decrease with sample size. At some point, the sample size will be sufficiently large that the MSE will approach zero for every parameter value. This is what the study of Fallin and Shorck (2000) demonstrated with a five locus system and sample sizes that are ≥ 50. An equivalent way to see whether a particular sample size is sufficient to reconstruct *all* haplotype frequencies with low MSE is to average the MSE over multiple simulated runs where the simulated data is based on the haplotype frequency vector with the highest entropy. This would be a vector of markers in linkage equilibrium, where the frequency of each marker allele is .5.

studies (i.e. large sample size with small numbers of loci).

2.5 Looking ahead: Extensibility of the EM Algorithm

The chapter has demonstrated that the EM-based approach remains one of the most reliable haplotype reconstruction algorithms for tightly linked markers. One additional attractive feature of the EM-based approach is that the algorithm can be extended to accommodate pooled data, where experimental design entails randomly partitioning sample of size N into K groups. EM-based models that estimate haplotype frequencies from pools are described in Yang et al. (2003) and Ito et al. (2003).

The appeal of DNA Pooling is that the procedure can be very cost efficient. It is conceptually evident that the standard EM-based haplotype reconstruction is equivalent to the case when $K = 1$. Cost decreases with the size of the pool, but larger pools provide less information regarding the latent haplotype data. Hence, the HW-EM algorithm is extensible to less informative genotype data than that provided by unrelated individuals.

This thesis will demonstrate that the EM algorithm can be also be extended to accommodate genotype data that is more informative than a sample of unrelated individuals, where the information is supplied by pedigree data. It will also examine the trade-off between cost and information. First, however, it is necessary to establish that such an algorithm is warranted, i.e. that existing haplotype reconstruction algorithms for pedigrees are not relevant. This is the focus of the next chapter.

Chapter 3

Haplotype Reconstruction for Pedigrees

3.1 Overview

Chapter 2 discussed haplotype reconstruction algorithms for unrelated individuals. These algorithms are stochastic since, in the absence of additional information (e.g. pedigree data), the chance of deterministically resolving phase is small, even when only a few loci are examined[1]. This chapter reviews existing approaches for haplotype reconstruction on pedigrees. Unlike haplotype reconstruction for unrelated individuals, there are two distinct classes of algorithms for pedigree data: stochastic and deterministic. These are reviewed in Sections 3.2 and 3.3 respectively.

Each of the two groups of algorithms rely on assumptions that are not appropriate for many experimental designs. Specifically, most stochastic methods assume markers are in Linkage Equilibrium (LE), while deterministic methods assume that pedigrees are sufficiently informative (large). Both assumptions were valid until recently, but because of high-resolution mapping methods, experimental design now favor smaller pedigrees and markers in high Linkage Disequilibrium (LD). Section 3.4 presents the results of a simulation study that demonstrates that neither paradigm is effective for small pedigrees and marker data that are tightly linked. The simulation study provides further justification for the model developed

[1]Recall that the only way a phase unknown genotype can be resolved deterministically is when the individual is heterozygous for at most one locus.

in the next chapter, which can accommodate LD and sparse pedigree data.

3.2 Stochastic Methods for Haplotype Reconstruction on Pedigrees

This section will discuss the Lander-Green algorithm, which serves as the core probability model for the most popular software packages for haplotype reconstruction on pedigrees: MERLIN(Abecasis et al., 2002), SIMWALK(Sobel and Lange, 1996) and GENEHUNTER(Kruglyak et al., 1996). The Lander-Green algorithm explicitly requires that markers are in LE. While stochastic haplotype reconstruction algorithms that accommodate LD have been proposed, these are inefficient, and will be discussed in Chapter 4.

3.2.1 The Lander-Green Algorithm

The Lander-Green algorithm was originally developed as a maximum likelihood method for estimating the recombination fraction (genetic distance) between markers. As with haplotype reconstruction, the observed data was multilocus genotype. For each marker genotype, the Lander-Green algorithm also defines an inheritance vector (equivalently an "IBD[2] vector"), which is regarded as the latent data.

The IBD vector species the meiotic outcome for each of the n non-founders in a pedigree. For a given pedigree member, the outcome is indexed as a bit, where 0 denotes a paternally derived allele and 1 denotes a maternally derived allele. A pedigree featuring n non-founders will require an inheritance vector of size 2^{2n} for each locus. This is why the Lander-Green algorithm is constrained by the number of members in the pedigree.

Once the inheritance vector is specified, information can be combined across the genome by using a Hidden Markov Model (HMM). Let G_i denotes the genotype for each pedigree member at locus i and I_i is the latent IBD vector at locus i. The three components of the likelihood that characterize an HMM are:

[2] Identical By Descent.

1. The transition probability, $P(I_i|I_{i-1})$. Note that if the states of I_i and I_{i-1} are the same, then no recombination has occurred, while if they differ than at least one recombinant event has occurred. Hence, this probability will be function of the recombination rate, r_i, between the two successive loci.

2. The prior probability, $P(I_0)$, which is assumed uniform over all possible IBD vector configurations, i.e. each state is assigned a probability of $\left(\frac{1}{2^{2n}}\right)$.

3. The emission probability, $P(G_i|I_i)$, which is defined as a product of the *marginal* allele frequencies for the founders. This reflects both the Hardy-Weinburg *and* the LE assumptions.

A graphical depiction of the likelihood for a parent-child trio appears in Figure 3.1. The IBD vectors are represented as a genetic descent graph, where each genotype as a pair of allele nodes, with the paternally derived allele listed first. Arcs connect the relevant parental allele node to a descendant child node.

3.2.2 Applications and Robustness of the Lander-Green Algorithm

As noted above, the Lander-Green algorithm was originally developed to infer recombination rates between loci. However, if genetic distances between loci are known, the Lander-Green formalism has many useful applications. For example, one of the most widely used applications of the Lander-Green algorithm is for Nonparametric Linkage Analysis (NPL). Specifically it can be used to test for an excess of IBD sharing among affected sib-pairs. Since it is straightforward to calculate $P(I|G)$, it is also possible to calculate an allele sharing statistic:

$$S(G) = \sum_I S(I)P(I|G),$$

where $S(I)$ is the number of IBD alleles shared by two affected sibs. This statistic can be compared to the expected number under no linkage.

Haplotype reconstruction is an obvious application. This can be achieved by inferring the

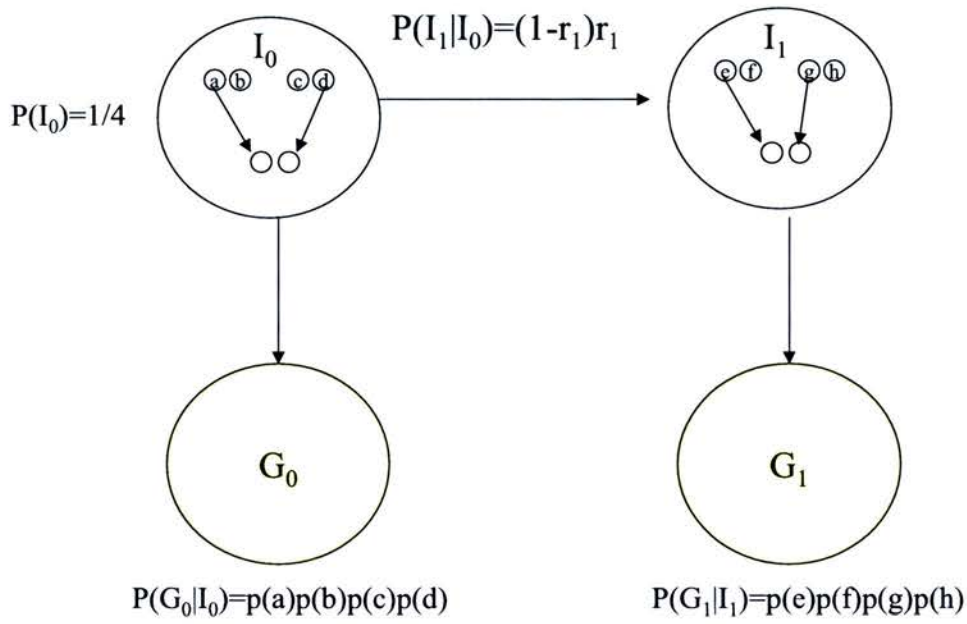


Figure 3.1: Lander-Green Representation of the Likelihood for a Parent-Child Trio at two loci. The observed data are the marker genotypes, G_i, for each of the i loci. The latent data are the corresponding IBD vectors, which are (0,1) and (1,1) for I_0 and I_1 respectively, are represented as a genetic descent graph. a-h refer to founder alleles, and p(a) - p(h) are their corresponding frequencies.

most likely sequence of latent data, e.g. using Viterbi sequence alignment. (The IBD vector at each locus is completely informative for the haplotype configuration of those loci).

With the advent of marker data in high LD, it is necessary to establish whether these Lander-Green applications are robust to violations of the LE assumption. Many studies have established that using markers in tight LD can severely bias results when using Lander-Green for NPL example discussed above(see, e.g. Abecasis and Wigginton, 2005).

Haplotype	Θ_{True}	Θ_{LE}
00	.5	.25
11	.5	.25
10	0	.25
01	0	.25

Table 3.1: Simple example illustrating dangers of assuming LE for markers in tight LD when reconstructing haplotypes

Evidence also exists that haplotype reconstruction accuracy can also be affected. The most widely cited study is by Schaid et al. (2002), which investigates a region for linkage using a case-control study. The objective of a case-control study is to compare haplotype frequencies between cases and controls. A significant difference is considered evidence for linkage. The region targeted by the study was spanned by three loci, all of which exhibited high pairwise LD. Cases were related, while controls were not. Haplotype frequencies were estimated for controls using the standard EM algorithm described in the previous chapter. Haplotype frequency estimates for cases were estimated using GENEHUNTER, which, as noted above, is a method that employs the Lander-Green framework. Haplotype frequencies were significantly different. However, when only unrelated cases were used and haplotype frequencies were estimated using the standard EM algorithm, no significant difference was detected. The LE assumption had resulted in false positive.

Section 3.4 presents the results of a simulation study that further establishes that Lander-Green based algorithms are inappropriate for tightly linked markers. However, Table 3.1 provides a simple example to motivate intuition. The Table describes the haplotype frequency profile, Θ_{True}, for two biallelic loci. Only two of the four haplotypes are segregating in the population, and each allele at each locus segregates with frequency 50%. If LE was assumed, the haplotype frequency of each haplotype would estimated as the product of its allelic frequencies. This would result in a radically different frequency profile, where each of the four haplotypes would be considered equally likely.

3.3 Deterministic Methods for Haplotype Reconstruction on Pedigrees

Many deterministic algorithms have been developed for haplotype reconstruction, and good review articles are available (see, e.g. Niu, 2004). All algorithms are characterized by the sequential and repeated application of a series of rules, and all algorithms suffer from two limitations. The first limitation is that these algorithms are only reliable when the pedigree is sufficiently informative. For example, rule-based approaches typically require that at least one parent is genotyped; the algorithm will simply not start in the event that even one progeny has untyped parents. More generally, if there are multiple haplotype configurations consistent with pedigree marker data, rule-based methods will either halt or return a single consistent configuration without assigning an appropriate degree of uncertainty.
The second limitation of deterministic methods is that no population information is used. To appreciate the importance of population data, consider the example where a parent-child trio has been genotyped at two loci and all three individuals are double heterozygotes. If the haplotype frequency profile is given by Table 3.2 then the haplotypes for each member in the pedigree could be reconstructed trivially.
It is clear that pedigrees need to be sufficiently informative for these algorithms to be effective. A threshold size has not been established, but full sibships of size 10 where both parents are genotyped was the sparsest pedigree considered by Nejati-Javaremi and Smith (1996). The method of Qian and Beckmann (2002), which is considered state-of-the-art, only examines pedigrees featuring at least 15 members. The review by Niu (2004) states the need to establish the limitations of rule-based methods for smaller, sparser, pedigrees. Section 3.4 provides further insight into the limitation of rule-based methods on sparse pedigree data.

3.4 Simulation Study: The Need for a New Paradigm

As stated previously, existing methods for haplotype reconstruction on pedigree data are not developed to accommodate tightly linked markers for small families. Instead, determinis-

tic methods can only accommodate larger pedigrees, while Lander-Green based methods can accommodate small pedigrees with markers in LE. As will be discussed in the following chapter, methods that can account for LD across loci are highly inefficient and, in practice, their utility is restricted by family size. In Chapter 4, a new model is developed that accommodate larger pedigrees a broader range of sparse pedigree structures. This section demonstrates the limitations of existing methods and the benefits of using this new paradigm.

Specifically, this section presents the results of a simulation study that is based on small independent half-sib pedigrees using empirically derived haplotype data. The half-sibships are paternal, where each dam gives birth to exactly one offspring. Scenarios where one or both parents are untyped are examined. The objective will be haplotype reconstruction for the sire. The simulation strategy is similar to those described in Chapter 5. Briefly, the simulation study entails: (1) specification of a haplotype frequency distribution for the parental population; (2) simulation of genotypes for independent half-sib pedigrees of various sizes; and (3) estimation of phase configurations for the sire.

The haplotype frequencies used are based on the African population featured in the study study by Hull et al. (2001). The study is typical of those being employed in fine-scale mapping analyses. The study examines six loci spanning a small (7.6 kb) region within a locus that is believed to contain a disease variant. As reported by the study, only 12 of the possible 64 haplotypes are segregating at these loci, which is indicative of high LD.

Three different algorithms are employed. The first is the "correct" one in that it utilizes population data and accommodates LD across loci. The second method is based on the Lander-Green paradigm in that it utilizes population data, but assume LE across loci. For both methods, the most likely haplotype configuration is selected.

A rule-based approach is also needed. Because standard rule-based methods require that at least one parent is genotyped, a genotype-elimination approach was implemented. The sire was successfully haplotyped if it could be assigned a unique haplotype configuration.

Results are presented Table 3.2. All three algorithms are comparable when both parents are typed and sibships are size 10. For sparser pedigrees, both stochastic methods outperform the rule-based approach. This is not surprising, since there is no need for stochastic

	SIRE PHASE CALL ACCURACY					
	Untyped Sires			Typed Sires		
Sib Size	**LD**	**LE**	**Rule**	**LD**	**LE**	**Rule**
1	0.07	0.00	0.00	0.91	0.66	0.42
2	0.14	0.02	0.00	0.98	0.78	0.58
3	0.30	0.09	0.02	1.00	0.85	0.70
4	0.52	0.20	0.04	1.00	0.88	0.80
5	0.65	0.25	0.10	1.00	0.95	0.90
10	0.92	0.50	0.30	1.00	1.00	1.00

Table 3.2: Percentage of Sire haplotypes that were accurately reconstructed using three different methods. Sires belong to half-sib pedigrees, and dams were untyped.

assessment if a pedigree can be resolved deterministically. What is most striking is how the model that correctly accounts for LD outperforms the model that incorrectly assumes markers are in LE. This distinction is most pronounced when both parents are untyped. Similar trends were observed when other haplotype frequencies were used. This simulation study provides further evidence that Lander-Green based algorithms are not robust to departures of the LE assumption.

Chapter 4

A Unified Model for Haplotype Reconstruction

This chapter presents a novel approach for reconstructing haplotypes for pedigree data featuring tightly linked markers. The approach is motivated by the key insight that a sample of unrelated individuals can be regarded as a collection of pedigrees, each of size one. Using graphical models, it is possible to extend the EM algorithm for unrelated individuals to accommodate more complex pedigree structures, such as paternal half-sibships. Section 4.1 provides an overview to the Chapter and introduces the key notation and assumptions that will be used throughout the chapter. Section 4.2 provides a qualitative description of graphical models and their application to haplotype reconstruction for paternal half-sibships. Section 4.3 provides a more mathematically detailed discussion of the inference algorithms used for haplotype reconstruction in paternal half-sibships.

4.1 Overview

Chapter 2 reviewed haplotype reconstruction algorithms for unrelated individuals. These algorithms were stochastic since, in the absence of additional information (e.g. pedigree data), the chance of deterministically resolving phase is small, even when only a few loci

are examined[1]. It was argued that that the popular EM-based method should be the default approach to haplotype reconstruction in population studies. One assumption made when discussing population studies in the abstract is that only samples of unrelated individuals will be ascertained (since these provide the most information about a population). In practice, however, pedigree data may be available for a population study, particularly in a fine-mapping analysis where pedigree data is used for the initial (coarse-mapping) linkage analysis. Two obvious problems arise if reconstruction algorithms for unrelated individuals are applied to pedigree data: first, the model is semantically wrong since members of the sample are no longer conditionally independent given the haplotype frequency profile; and second, valuable information that can help resolve phase ambiguity is ignored.

The objective of this chapter is to extend the EM-based approach to accommodate paternal half-sib pedigrees, which is a commonly encountered structure for many species of livestock. Specifically, a model is presented to conduct exact inference on arbitrary large half-sibships that explicitly account for LD across loci. Loci are assumed tightly linked, as would be expected in a high resolution mapping study. This assumption has two important consequences: haplotypes are likely to nonrecombinant between two successive generations and existing stochastic haplotype reconstruction algorithms (which, as discussed in Chapter 3.1, are not robust to the violations of the LE assumption) are inappropriate.

As noted above, breeding schemes featuring paternal half-sibs are common for many species of livestock. This chapter considers breeding schemes where each pregnancy will result in the birth of a single offspring (e.g. dairy cattle). The consequence of this assumption is that pedigree structures are too sparse for the deterministic algorithms discussed in Chapter 3.1. The most widely cited stochastic model for LD-based haplotype reconstruction on sparse pedigrees is the method of Rohde and Fuerst (2001). Their inference algorithm, which is developed for full sibs, is extremely inefficient (it cannot accommodate more than a few sibs); adapting their methodology to the half-sib pedigree structure is not appropriate. The reason why their algorithm is so inefficient is because it requires exhaustive enumeration of all haplotype configurations consistent with genotype in the pedigree. Zhang et al. (2005)

[1]Recall that the only way a phase unknown genotype can be resolved deterministically is when the individual is heterozygous for at most one locus.

improve on the method of Rohde and Fuerst (2001) by introducing a set of rules that efficiently eliminate inconsistent haplotype configurations in the pedigree. However, when there is missing data, the number of haplotype configurations will not be substantially reduced and the method will suffer from the same problems as the method of Rohde and Fuerst (2001). As an alternative, O'Connell (2000) alludes to the possibility of modifying algorithms developed in the context of human linkage analysis to create efficient LD-based haplotype reconstruction algorithms, but this idea was not fully developed.

This chapter proposes an alternative framework that is powerful, elegant and (most importantly) robust to achieve this objective: probabilistic graphical models. This framework entails specifying the relevant joint probability distribution as a graph, and then manipulating the graphical structure to facilitate inference. Graphical models should be regarded a tool to help facilitate principled probabilistic inference for difficult problems[2].

At the time this thesis was being undertaken, graphical models were gaining recognition as being useful for formulating and solving problems in genetics. Lauritzen and Sheehan (2003) provide an generic overview of the graphical model paradigm, with some applications to simple linkage analysis and QTL detection. As noted by Lauritzen and Sheehan (2003) graphical models are a natural way to derive the "peeling" algorithms (Elston and Stewart, 1971; Cannings et al., 1978) that had been developed to conduct efficient inference in pedigrees in the context of linkage analysis.

The popular statistical package SUPERLINK (Fishelson and Geiger, 2002) conducts efficient linkage analysis and was developed using graphical models. The software was modified to accommodate haplotype reconstruction (Fishelson et al., 2005) at the same time that Schouten et al. (2005) was published. The two approaches (including the likelihood function, graphical representation and inference algorithms) are different. A fundamental difference is that Fishelson et al. (2005) assumes that markers are in LE while Schouten et al. (2005) assume markers are in tight LD. As noted in the previous chapter, LE should not be assumed when reconstructing haplotypes with markers in tight LD.

After introducing some additional notation, the chapter provides a qualitative description

[2] A suitable analogy are roadmaps, which help us visualize various route options (and their consequences) as well as offer associated algorithms (e.g. greedy search) to solve challenging optimization problems (e.g. the travelling salesman problem).

of graphical models and their application to haplotype reconstruction for paternal half-sibships. Important insights into the haplotype reconstruction problem that were gained through by using graphical models are also presented. The chapter concludes by providing a more mathematically detailed discussion of the inference algorithms used for haplotype reconstruction in paternal half-sibships.

4.1.1 Notation and Assumptions

It is useful to briefly review the EM algorithm for unrelated individuals that was discussed in Chapter 2. Recall that each iteration of the EM algorithm requires calculating the expected log-likelihood of the "complete" data, which consists of the observed phase-unknown genotypes ($\mathbf{y}$) and the analogous latent phase configurations ($\mathbf{z}$) :

$$E_{p(\mathbf{z}|\mathbf{y},\tilde{\Theta})} \log[p(\mathbf{y},\mathbf{z}|\Theta)] \;=\; \sum_{i=1}^{N}\sum_{j=1}^{M} E_{p(z_i|y_i,\tilde{\Theta})} n_{ij} \log \theta_j + \text{Constant}, \tag{4.1}$$

where $\tilde{\Theta}$ denotes the current estimate of Θ and n_{ij} refers to the number of times haplotype j appears in the phase configuration of individual i.
We now consider a sample of half-sib pedigrees. We assume that the parental generation is in HWE. Below are five additional assumptions[3]:

1. Sires can be mated to multiple dams.
2. Dams can be mated with exactly one sire.
3. Each dam can have exactly one progeny.
4. Relationships in the sample are known with certainty.
5. Genotypes are available for all progeny, but are unavailable for sires and dams.

Formally, we consider the case where members of $\mathbf{y}$ may be related through one of $P \leq N$ sires. If $P < N$, then the likelihood in (4.1) is no longer valid since the summation over

[3]Assumptions 1-4 would be valid for dairy cattle. (Assumption 5 can be valid for any species). Section 4.3.4 demonstrates how this model can be extended to accommodate typed parents and full sibs.

N only follows from assuming each animal is unrelated. Clearly, the likelihood must be revised to accommodate pedigree structure (as well as any parental genotype information). It is first necessary to introduce notation describing parental marker data. Let $\mathbf{s} = s_1 \ldots s_P$ denote the unobserved phase configurations for the P sires. y_{ki} will be used to signify that sire i is the parent of animal k, while $\mathbf{y}_{.i}$ denotes the set of sampled genotypes that are related through sire i. Thus, $\eta_i = |\mathbf{y}_{.i}|$ specifies the size of this sibship.

Given the assumption that each dam will have exactly one progeny, η_i refers to the number of progeny *and* the number of dams affiliated with sire i. Similarly, let d_{ki} denote the unobserved phase configuration for the dam of animal k while $\mathbf{d}_{.i}$ is the set of phases for all dams that were mating with sire i. Without loss of generality, we will assume the parents are untyped. We also make the assumption that each dam can have exactly one offspring. The complete data can be then specified as a collection of P half-sibships:

$$(\mathbf{s}, \mathbf{d}, \mathbf{y}) = \bigcup_{i=1}^{P} (s_i, \mathbf{d}_{.i}, \mathbf{y}_{.i}). \tag{4.2}$$

4.2 The Graphical Model Paradigm

This section provides a brief description of the graph-theoretic algorithms that will facilitate inference in the haplotype reconstruction model for paternal half-sibships. Several key insights into the problem domain were gained using the graphical model paradigm, and these are highlighted. As noted above, inference using graphical models entails representing the relevant joint probability distribution as a graph and then using associated algorithms to conduct inference. Broadly, the graphical model formalism provides powerful tools to facilitate model specification, visualization, and inference.

The two most common forms of graphical models are based on directed acyclic graphs (DAGs)[4] and undirected graphs. In both forms, random variables are represented as nodes and the the joint distribution is expressed as the product of local functions that are defined over connected subset of nodes. However, the algorithms associated with model specifica-

[4]Also known as Bayesian Networks.

tion and inference are different[5]. When the objective of inference is to obtain the marginal distributions for all of the latent variables in the joint distribution (as is the case with haplotype reconstruction), it is preferable to work with a Junction Tree, which is a data structure that is based on undirected graphs[6]. However, when specifying distributions describing pedigrees, it is natural to using algorithms associated with DAGs. Fortunately, it is not necessary to choose which form to use since there is a process for creating and conducting inference on a Junction Tree that respects the conditional independence statements of a previously stated DAG. This is the procedure that was used to generate the results presented in the previous section, and which is described below.

4.2.1 Model Specification and Visualization

The process of model specification using DAGs follows a process that mirrors model specification using the chain rule of probability: An ordering (topology) of the variables (nodes) is proposed; each random variable is introduced onto the graph (as a node); and a directed arc is drawn from the existing nodes that have a direct influence on that variable. For each node, the conditional probability distribution of that node given its parents is specified. As with the chain rule, the joint distribution is then specified as the product of these "pruned" conditional probability distributions.

In the context of the half-sib pedigree, the topology is based on a descending ordering of variables according to generation, i.e. $\{\mathbf{s}, \mathbf{d}, \mathbf{z}, \mathbf{y}\}$. The DAG is depicted on the right panel of Figure 4.1. Shaded nodes denote observed variables[7]. The nested blocks (plates) provide a compact way to represent the replication that occurs in the experiments (i.e. they have the same semantics as the product symbols in a conventional mathematical description of the likelihood).

[5]It should be noted that each graph makes different assertions about conditional independencies, and there there are certain probability distributions that can only be expressed by one of the graphical forms.

[6]This is a key reason why "undirected graphs play a crucial role in solving inference and learning problems efficiently, even for models whose definition is based on a directed graph"(Jordan, 2003).

[7]Since we are using a frequentist approach to inference, the haplotype frequency profile Θ cannot be regarded as a node since it is not a random quantity. It is used to underscore that joint distribution is a function of unknown parameter.

The joint probability distribution specified by a DAG is the product of local functions defined on each node. As noted above, the location function of each node is the conditional probability distribution of that node given its parents. The complete data log-likelihood specified by the DAG in Figure 4.1 is therefore:

$$\begin{aligned}\log L_C(\Theta) &= \log\{p(\mathbf{s},\mathbf{d},\mathbf{y}|\Theta)\} \\ &= \sum_{i=1}^{P}\log p(s_i|\Theta)+\sum_{i=1}^{P}\sum_{k=1}^{\eta_i}\log p(d_{ki}|\Theta)+\sum_{i=1}^{P}\sum_{k=1}^{\eta_i}\log p(y_{ki}|s_i,d_{ki}),\end{aligned} \tag{4.3}$$

where $p(s_i|\Theta)$ and $p(d_{ki}|\Theta)$ follow the same distribution given by equation (2.1), and

$$p(y_{ki}|s_i,d_{ki}) = \sum_{z\in\mathbf{z}(y_{ki})} p(z|s_i,d_{ki})p(y_{ki}|z). \tag{4.4}$$

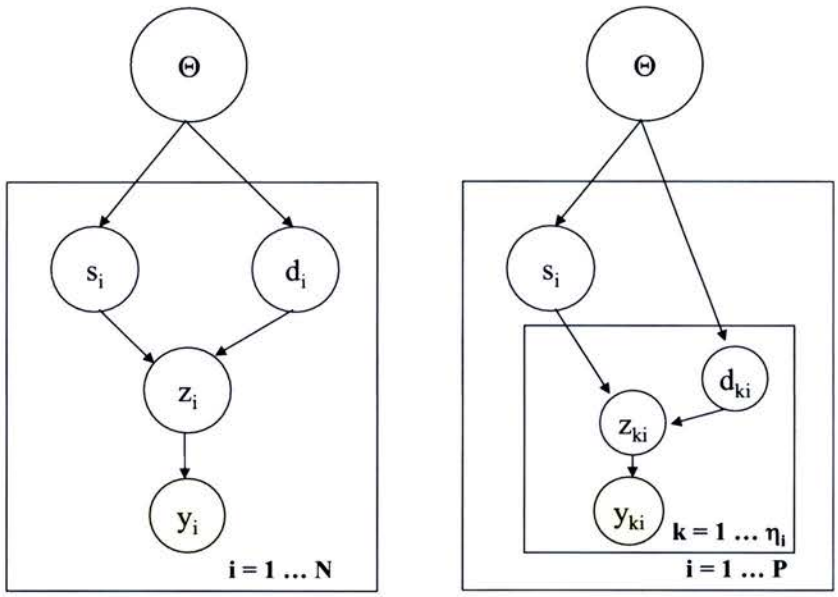
There are two important benefits to visualizing the likelihood using graphical models that directly impacted the course of this research. First, note that if a dam has exactly one offspring per mating (as would be the case for dairy cattle) and both parents are untyped, then this model will yield the same results as one designed for unrelated individuals. Hence this model can be regarded as a unified approach to haplotype reconstruction. This observation has important benefits which will be seen in the next chapter. The graphical representation also reveals that this model has uses beyond haplotype reconstruction. One example, which is explored in the next chapter, is that it is well-suited for assessing optimal resource allocation with fixed resources.

4.2.2 Inference

As will be discussed in more detail in the next section, maximum likelihood estimation of Θ will involve efficient calculation of $p(s_i|\Theta)$ and $p(d_{ki}|\Theta)$. This section provides the derivation of $p(s_i|\Theta)$ and $p(d_{ki}|\Theta)$ using the graphical model framework. Equivalently, we describe an efficient process for estimating the marginal distributions for all latent variables (i.e. phase) in the pedigree.

The inference algorithm that is associated with DAGs is Variable Elimination. Specifically,

Figure 4.1: DAG for the joint distributions relevant to haplotype reconstruction for unrelated individuals (left) and half-sib pedigrees (right).



obtaining the marginal distributions for the a set of variables entails summing over all remaining latent variables. The key idea is to push the sums into the factorized distribution as far as possible (using the distributive law) and then perform the sums recursively. This, of course, is precisely the idea behind the Elston and Stewart (1971) peeling algorithm (variable elimination). Peeling is appropriate in the context of classical linkage analysis since the objective is to marginalize all of the latent variables (which allows the recombination rate to estimated). The marginal distributions of the latent variables were not required. Here, the situation is reversed: the recombination rate is assigned a value of zero and the marginal distributions are needed.

To obtain the marginal distribution for each of the latent variables within the DAG framework, it would be necessary to apply the peeling algorithm repeatedly. An efficient alterna-

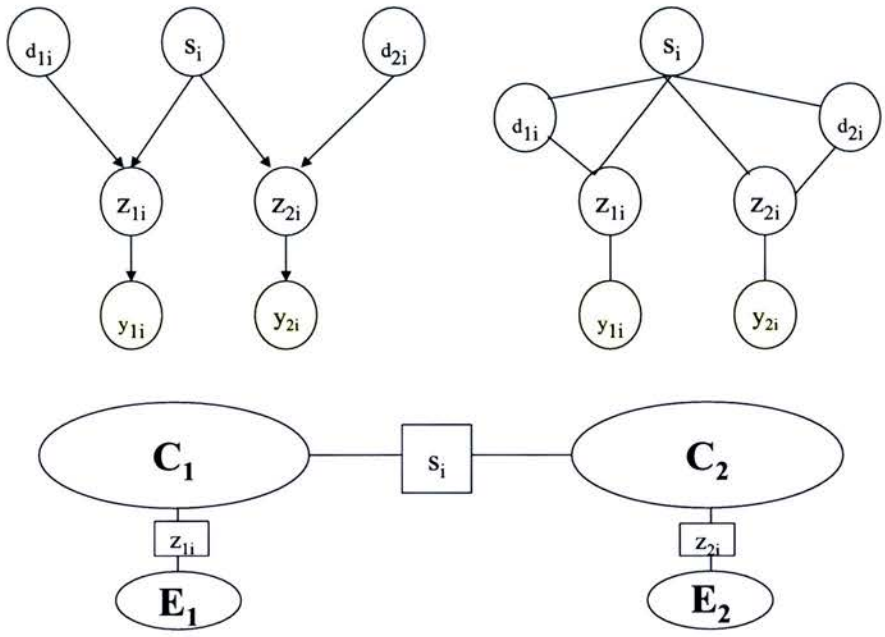


Figure 4.2: Deriving the Junction Tree for a Half-Sib Pedigree, where $\eta_i = 2$ From top left: (1) the original DAG; (2) the moralized (undirected) graph and (3) the Junction Tree. The cliques, which are read off the moralized graph, are defined as follows: $C_1 = \{s_i, d_{1i}, z_{1i}\}, C_2 = \{s_i, d_{2i}, z_{2i}\}, E_1 = \{z_{1i}, y_{1i}\}$ and $E_2 = \{z_{2i}, y_{2i}\}$.

tive is to employ the Junction Tree Algorithm, which is a systematic method for efficiently computing all marginal probabilities. This is an algorithm that is based on *un*directed graphs. To use this algorithm, one must first convert the DAG into an appropriate Junction Tree. The first step is to replace the directed arcs with undirected arcs and add links to connect nodes with common descendants. This process is called moralization and results in an undirected graph that respects the conditional independence statements in the DAG. In addition to moralization, it is also necessary to add arcs to eliminate chordless cycles. This process is referred to as triangulation, and will not be needed for half-sib pedigrees. convert the triangulated graph into a Junction Tree. This involves first identifying all the maximal cliques in the moralized graph (i.e. all fully connected subsets of nodes that cannot

be extended without losing the property of being fully connected); and finally constructing a maximal-spanning tree (e.g. by using Kruskal's algorithm) out of the cliques. Additional nodes (separators) are introduced between each pair of cliques and contain variables that are common to the two cliques.

The process of deriving the Junction tree for a pedigree with two half sibs, i.e. $\eta_i = 2$ is sketched in Figure 4.2 (to minimize notational clutter explicit references to Θ are dropped). Figure 4.3 depicts the Junction Tree for the general case, i.e. for a sibship of size η_i.

Once the Junction tree has been created, one defines a potential function over each clique by assigning each factor in the DAG to any one clique that features all relevant variables. The potential function for that clique is the product over each of these factors.

For the sibship in Figure 4.3, it is clear that the potential for E_k, Ψ_{E_k}, must be:

$$\Psi_{E_k} = p(y_{ik}|z_{ik}) \qquad k = 1 \ldots \eta_i$$

One option for the clique potentials for C_k are:

$$\Psi_{C_k} = \begin{cases} p(s_i)p(d_{ik})p(z_{ik}|s_i, d_{ik}) & k = 1 \\ p(d_{ik})p(z_{ik}|s_i, d_{ik}) & k = 2 \ldots \eta_i \end{cases}$$

The separators are also assigned potential functions, and these are initialized to unity. Once the Junction Tree has been constructed, a straightforward, generic process for inference is employed that, after a single run, guarantees that the expression for all nodes will contain marginals of the relevant variables. The algorithm is applied in the next section, but is outlined here:

First, designate a root of the Junction Tree and update the separators and clique potentials from the leaf of the tree to the root as follows:

$$\phi_S^* = \sum_{V \setminus S} \psi_V \tag{4.5}$$

$$\psi_W^* = \frac{\phi_S^*}{\phi_S} \psi_W. \tag{4.6}$$

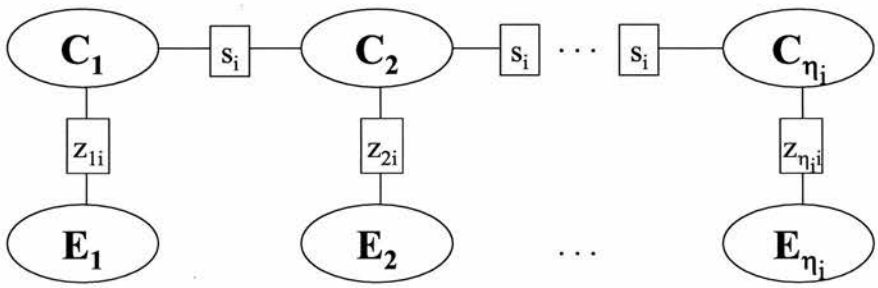


Figure 4.3: The Junction Tree for a half-sib pedigree of size η_i. The cliques are defined as: $C_k = \{s_i, d_{ki}, z_{ki}\}$, $E_k = \{z_{ki}, y_{ki}\}$.

Next, update the clique potentials from the root of the tree to the leaf as follows:

$$\phi_S^{**} = \sum_{V \setminus S} \psi_V^* \tag{4.7}$$

$$\psi_W^{**} = \frac{\phi_S^{**}}{\phi_S^*} \psi_W^*. \tag{4.8}$$

Once the algorithm has been completed, then for each variable x of interest, identify a clique, C containing x. $p(x)$ is computed as follows:

$$p(x) = \sum_{C \setminus x} \psi_C^{**}.$$

4.3 An EM Algorithm for Outbred Half-Sib Pedigrees

The section provides a straightforward mathematical description of an EM algorithm for outbred half-sib pedigrees. The graph-theoretic algorithms that facilitated the derivation of this algorithm are presented in Section 4.2. The section is organized as follows: first, the appropriate likelihood is specified; second, an efficient inference algorithm is described; and, finally, an the complexity of this algorithm is evaluated.

4.3.1 Specifying the Complete-Data Log-Likelihood

Each iteration of the EM algorithm entails calculating the expectation of (4.3) with respect to $p(\mathbf{s},\mathbf{d}|\mathbf{y},\tilde{\Theta})$. Note that the third term in (4.3) does not depend on Θ, and can be regarded as a constant term in the context of this analysis. The expectation of the complete data (omitting this constant term) can be expressed as:

$$\sum_{i=1}^{P}\sum_{j=1}^{M} E_{p(s_i|\tilde{\Theta},\mathbf{y})}\left[n_{ij}\right]\log\theta_j \;+\; \sum_{i=1}^{P}\sum_{k=1}^{\eta_i}\sum_{j=1}^{M} E_{p(d_{ki}|\tilde{\Theta},\mathbf{y})}\left[n_{kij}\right]\log\theta_j, \tag{4.9}$$

where n_{kij} refers to the number of times haplotype j appears in the kth dam of sire i.
This expression should be contrasted with (4.1). The key distinction involves calculating the distribution of the latent phase configurations. For unrelated individuals, calculating $p(\mathbf{z}|\mathbf{y},\tilde{\Theta})$ is straightforward (and provides a clear bound on the tractability of inference). The analogous challenge for this model is to calculate $p(s_i|\tilde{\Theta})$ and $p(d_{ki}|\tilde{\Theta})$ efficiently, which is less straightforward.

4.3.2 Calculating the Marginal Distributions of the Latent Data

To minimize notational clutter explicit references to $\tilde{\Theta}$ are dropped; and it should be understood that this information is given. Furthermore, since the relevant distribution is for a specific sire or dam, we drop any index that refers to the sire, the sire's dams or offspring. Finally, we will assume the sibship is size K (i.e. $\eta_i = K$).
Calculating the Distribution of the Sire Phase: The objective can therefore be written as

$$p(s|\tilde{\Theta},\mathbf{y}) \doteq p(s|\mathbf{y}) \;=\; \frac{\sum_{\mathbf{d}} p(s,\mathbf{d},\mathbf{y})}{p(\mathbf{y})}. \tag{4.10}$$

The joint distribution is expressed as a telescopic sum, which was the key insight of Elston and Stewart (1971):

$$p(\mathbf{y},s) = p(s)\prod_{k=1}^{K}\left\{\sum_{d_k} p(y_k|s,d_k)p(d_k)\right\}, \tag{4.11}$$

where

$$p(y_k|s,d_k) = \sum_{z \in \mathbf{z}(y_k)} p(z|s,d_k)p(y_k|z). \tag{4.12}$$

Peeling the jth family first entails calculating

$$p(s,\mathbf{y}_j) = \sum_{d_j} p(s,\mathbf{y}_{j-1})p(y_j|s,d_j)p(d_j) \tag{4.13}$$

where $\mathbf{y}_{j-1} = y_0, y_1 \ldots y_{j-1}$ and $y_0 = \emptyset$. The likelihood is then updated to

$$p(\mathbf{y}) = p(s,\mathbf{y}_j) \prod_{k=j+1}^{K} \left\{ \sum_{d_k} p(y_k|s,d_k)p(d_k) \right\}. \tag{4.14}$$

Each family is iteratively peeled and, after the final family has been peeled, the resulting expression, $p(\mathbf{y},s)$, can be used to calculate $p(s|\mathbf{y})$. $p(\mathbf{y})$ can then be obtained by summing out s.

Calculating the Posterior for the Dam: First, the objective can be rewritten as

$$\begin{aligned} p(d_j|\mathbf{y}) &= \sum_s p(s,d_j|\mathbf{y}) \\ &= \frac{1}{p(\mathbf{y})} \sum_s p(s,d_j,\mathbf{y}) \\ &= \frac{1}{p(\mathbf{y})} \sum_s p(s,\mathbf{y}_{-j})p(y_j|d_j,s)p(d_j), \end{aligned} \tag{4.15}$$

where efficient calculation of $p(\mathbf{y})$ is given above and the definition of $p(s,\mathbf{y}_{-j})$ is given by:

$$p(s,\mathbf{y}_{-j}) = p(s,y_0,y_1 \ldots y_{j-1},y_{j+1},\ldots y_K).$$

To calculate $p(s,\mathbf{y}_{-j})$, it is necessary to store each of the K expressions given by equation (4.13), i.e.

$$p(s,\mathbf{y}_j) \qquad j = 0 \ldots K. \tag{4.16}$$

These are sufficient to calculate $p(s,\mathbf{y}_{-j})$ since:

$$\begin{aligned}\left[\frac{p(s,\mathbf{y})}{p(s,\mathbf{y}_j)}\right]p(s,\mathbf{y}_{j-1}) &= \frac{\prod_{k=1}^{K} p(y_k|s)}{\prod_{k=1}^{j} p(y_k|s)}p(s,\mathbf{y}_{j-1}) \\ &= \prod_{k=j+1}^{K} p(y_k|s)p(s,\mathbf{y}_{j-1}) \\ &= p(y_{j+1}\cdots y_K|s)p(s,\mathbf{y}_{j-1}) \\ &= p(s,\mathbf{y}_{-j}). \end{aligned} \tag{4.17}$$

4.3.3 Evaluating the Complexity of the Algorithm

The crucial property of this method is that complexity scales linearly with the size of the sibship[8]. From equations (4.12) and (4.17), it can be seen that the complexity of the phase distribution for both sire and dam is dominated by the expression $p(y_k|s,d_k)$, which is cubic in the number of phase configurations. Hence the complexity is $O(KM^3)$

This is still two orders of magnitude worse than the complexity for unrelated individuals. Fortunately, there are ways to reduce the complexity further for both sire and dam. In the context of haplotype reconstruction, the corresponding reduction in computational resources can be substantial.

Quadratic complexity can be achieved by evaluating $p(y_k|s)$ and $p(y_k|d_k)$ directly. Consider first evaluating the relevant expression for the sire where the genotype of the dam is unknown. Rather than summing over all the dam configurations, as suggested by (4.11), consider directly evaluating

$$p(\mathbf{y},s) = p(s)\prod_{k=1}^{K} p(y_k|s). \tag{4.18}$$

A given set of phase configurations, $z \in \mathbf{z}(y_k)$, can be expressed as

$$p(z = h_i, h_j | s = h_k, h_l). \tag{4.19}$$

[8]By contrast, the complexity that would result from adopting the method described by Rohde and Fuerst (2001) would be exponential with respect to the size of the sibship.

Let $t_1(s)$ denote the event that the sire transmitted h_k and let $t_2(s)$ denote the event that the sire transmitted h_l. Then (4.19) can be written as:

$$\begin{aligned} p(z=h_i,h_j|s=h_k,h_l) &= p(z,t_1(s))+p(z,t_2(s)) && (4.20)\\ &= p(z|t_1(s))p(t_1(s))+p(z,t_2(s))p(t_2(s)) && (4.21)\end{aligned}$$

where $p(t_1(s)) = p(t_2(s)) = 1/2$. Importantly, $p(z|t_1(s))$ and $p(z|t_2(s))$ can be calculated directly from $\tilde{\Theta}$, for example:

$$p(z|t_1(s)) = \begin{cases} \tilde{\theta}_j & k=i \\ \tilde{\theta}_i & k=j \\ 0 & \text{otherwise.} \end{cases} \quad (4.22)$$

When the genotype of the dam is known, the probabilities in (4.22) can be calculated from the distribution $p(d|\tilde{\Theta})$.

4.3.4 Extensions to the Graphical Model

This section discuss two straightforward extension to the half-sib model that was derived in this chapter. The first extension accommodates parental genotypes. The second extension accommodates full-sibs. It is important to realize that the probability model used in these two extensions is exactly the same as that used in outbred half-sib pedigrees, i.e. the probability model reflects a random mating population in HWE.

4.3.4.1 Introducing Additional Evidence

In the previous sections, we have assumed that parental genotype data is unavailable. We can augment the graph in Figure 4.1 to include parental genotypes as shown in Figure 4.4. The likelihood is therefore

$$\begin{aligned} \log L_C(\Theta) &= \log\left\{p(\mathbf{s},\mathbf{d},\mathbf{y},\mathbf{y}^{\mathbf{s}},\mathbf{y}^{\mathbf{d}}|\Theta)\right\} \\ &= \sum_{i=1}^{P}\log p(s_i|\Theta)p(y_i^s|s_i)+\sum_{i=1}^{P}\sum_{k=1}^{\eta_i}\log p(d_{ki}|\Theta)p(y_{ki}^d|d_{ki})+\sum_{i=1}^{P}\sum_{k=1}^{\eta_i}\log p(y_{ki}|s_i,d_{ki}). \end{aligned} \quad (4.23)$$

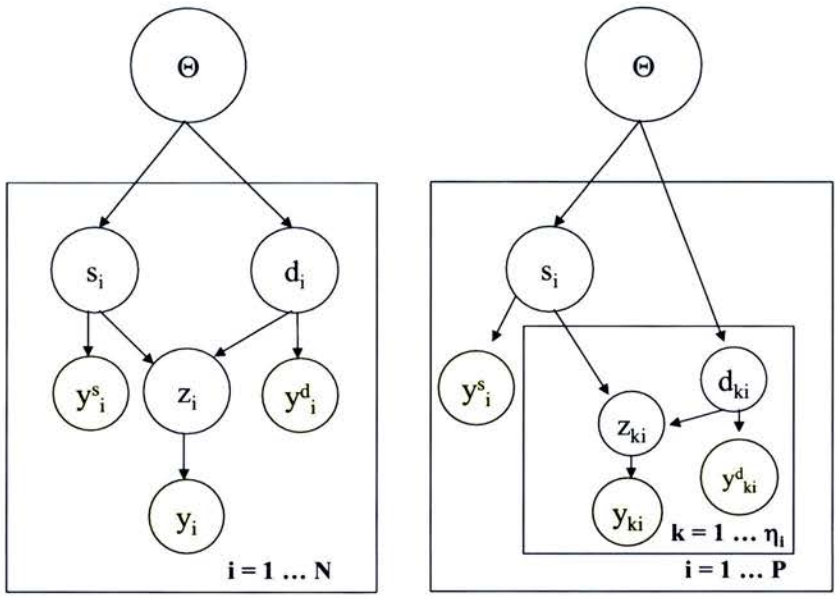


Figure 4.4: DAG for the joint distributions relevant to haplotype reconstruction for unrelated individuals (left) and half-sib pedigrees (right) when parental genotypes are included.

The introduction of parental genotypes simply eliminates some of the phase options that had been exhaustively enumerated in the parents when parental genotype data was unavailable.

4.3.4.2 Introducing Full Sibs

Full sibs can easily be accommodated, and the basic data structure is shown in Figure 4.5. Note that the junction tree algorithm, specified by equations 4.5 through 4.8, is exactly the same. The only difference is the dimension of the separator potential, which is now quadratic in the number of phase configurations.

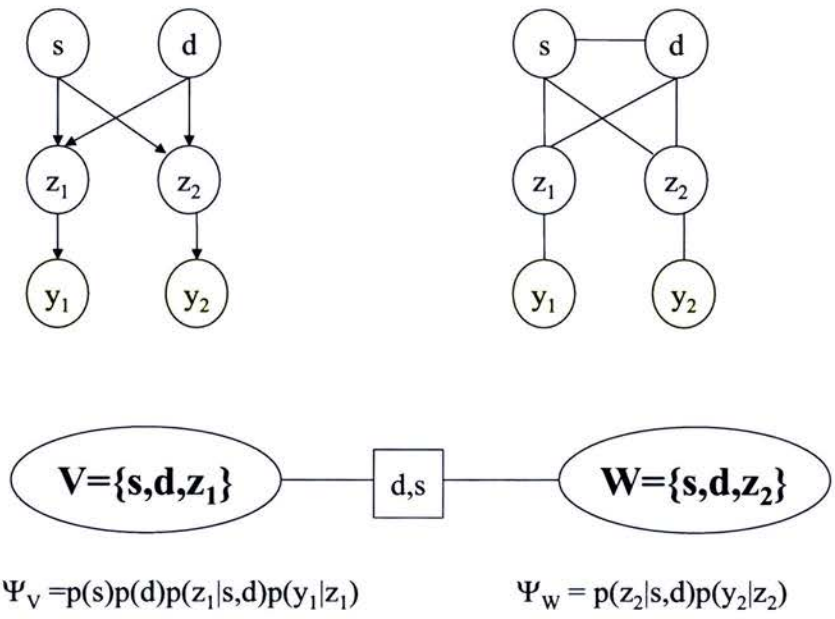


Figure 4.5: Deriving the Junction Tree for a Full-Sib Pedigree. From top left: The original DAG; the moralized (undirected) graph; and, the Junction Tree.

Chapter 5

Simulation Studies

This chapter uses the model developed in Chapter 4 to conduct two important simulation studies. The first study examines the effectiveness of using family data to improve accuracy for both haplotype frequency estimates and phase assignments. The second study investigates the consequences failing to account for relatedness from samples drawn from small hierarchical populations. The results from the two studies demonstrate the crucial role that the haplotype frequency profile (which is defined by patterns of LD) can have in determining the reliability of inference. Treating related individuals as unrelated in this context is common practice (see, e.g. the "LD only" analyses in Lee and van der Werf, 2004; Meuwissen et al., 2002) and we show that this can adversely impact haplotype frequency accuracy. The chapter concludes with a discussion of the practical implications of these results for experimental designs in population studies.

5.1 Overview

As stated previously, there are two conventional approaches to inferring haplotypes from population data, both with potential drawbacks. One approach is to use family data, which may be able to deterministically resolve phase for genotypes featuring multiple heterozygous loci. A serious drawback is that ascertaining this information can be costly[1]. A second

[1] Also, as we showed in Chapter 3, there is no guarantee that family data will resolve phase ambiguity.

approach, which was discussed in Chapter 2, is to infer haplotypes directly from population data. A problem with this approach is that there may be considerable uncertainty associated with the inferred haplotypes.

Many research groups want to know how much family data (if any) should be used to facilitate haplotype reconstruction in a population study. Section 5.2 presents a simulation study that evaluates how changes in family information can affect the accuracy of haplotype frequency estimates and phase reconstruction. Results from this simulation study suggest that treating related individuals as unrelated can significantly impact the quality of inference. In Section 5.3, these results are explored in a more realistic simulation environment. Specifically, the second study investigates the consequences failing to accommodate relatedness in samples drawn from small hierarchical populations. The two simulation studies are complementary and their relation to each other can be summarized as follows: in the first simulation study, the sibsize for each sample is fixed at x, while in the second study the expected sibsize for each sample is x. Both studies reveal the central role of the true haplotype frequency distribution in the overall quality of inference. 5.4.

5.2 Simulation Study: The Effectiveness of Pedigree Data in Haplotype Reconstruction

A simulation study based on independent half-sib pedigrees using empirically derived haplotype data is used to examine the effectiveness of family data in haplotype reconstruction. The simulation strategy is divided into the following three steps: (1) specification of a haplotype frequency distribution for the parental population; (2) simulation of genotypes for independent half-sib pedigrees; and (3) estimation of haplotype frequencies and phase configurations using different categories of missing data and different assumptions about relatedness.

5.2.1 Specifying the Haplotype Frequency Distributions

The most important parameter in the simulation study is the parental haplotype frequency distribution. The simulation study uses three empirically-derived haplotype frequency distributions. The first two frequency profiles, $APOE_1$ and $APOE_2$, were provided by Fallin et al. (2001) and correspond to two sets of marker data for a control group used in an association study for Alzheimer's Disease. The third data set, $IL8_E$ was presented by Hull et al. (2001) and corresponds to haplotype frequency estimates of a European sample for six biallelic loci spanning a 7.6 kb region within the $IL8$ locus.

One of the central results from the simulation study is that the expected accuracy for any estimate will be different for each of the population frequency distributions. It will be useful to identify relevant summary statistics that capture the relative performance that can be expected for random samples from each of the populations.

Qin et al. (2002) demonstrate that the variance for each EM-based estimate of θ_i can be expressed as the sum of two components: the first component reflects the variance of θ_i if phase configurations are observed, while the second component reflects the loss of information because of unknown phase configurations.

When phase is known, the uncertainty associated with the distribution, Θ, is best described by the entropy, i.e. $-\sum_i \theta_i \log \theta_i$. A more biologically relevant metric that measures the uniformity of the frequency distribution is the gene diversity, $1 - \sum_{i=1}^{M} \theta_i^2$.

To describe the additional uncertainty from the unknown phase configurations, an appropriate metric is the expected error rate using most likely phase configuration. Consider a given phase-unknown genotype, y. The probability that the most likely phase configuration is the correct one is given by $\max p(z|y,\Theta)$. A measure of the uncertainty from not knowing phase for this genotype is $1 - \max p(z|y,\Theta)$. The expectation of incorrectly assigning phase for a random sample is therefore:

$$E(\varepsilon|\Theta) = \sum_y p(y|\Theta)\left[1 - \max p(z|y,\Theta)\right]. \tag{5.1}$$

Appreciating the relevance of equation (5.1) in the context of accurate EM-based phase reconstruction cannot be overstated. The expression describes the number of incorrect phase

	$APOE_1$	$APOE_2$	$IL8_E$
Number of Loci	4	4	6
Number of Haplotypes	13	10	9
Ambiguous Genotypes[1]	0.52	0.41	0.53
Gene Diversity	0.86	0.8	0.56
Frequency-Known Error Rate[2]	0.18	0.12	0.0003

[1] The expected proportion of a population sample that is heterozygous for at least two loci.
[2] The expected proportion of phase configurations that will be incorrectly resolved in a population sample when the most likely phase criterion is used and haplotype frequencies are known.

Table 5.1: Summary statistics for haplotype frequency estimates used in data analysis.

assignments that is expected in a population sample when the most likely phase configuration is used and haplotype frequencies are known. It can therefore be considered a lower bound on the number of errors that are calculated from haplotype frequencies inferred by the EM algorithm.

These two statistics are presented in Table 5.1. If the population haplotype frequency for $APOE_1$ were known with certainty, one would expect to get no greater than 82% of the sample correct if the most likely phase criterion is used. By contrast, phase assignment using the most likely phase criterion would be virtually error free for population samples generated from the $IL8_E$ distribution, even though the expected number of ambiguous genotypes (i.e. genotypes with two or more heterozygous loci) is similar to the $APOE_1$. Although there are multiple phase configurations that can, in theory, resolve an ambiguous genotype sampled from $IL8_E$, the vast majority of these will feature at least one haplotype that does not actually segregate in the population. This demonstrates that family data may be unnecessary for accurate phase reconstruction, even when a sample features many ambiguous genotypes.

5.2.2 Simulating the Data

The number of sampled individuals is fixed at 100 and sib-sizes are fixed at 1,2,5,10 and 25. The categories of family information that are used with each sample when reconstructing haplotypes are given in Table 5.2. Since family sizes are exact, results for samples

Category	Description
UN	No Information - All Animals Assumed Unrelated
P	Pedigree Structure Only (Parents Untyped)
PS	Pedigree Structure + Sire Genotype
PSD	Pedigree Structure + Sire and Dam Genotypes

Table 5.2: Categories of family data that can be included with a sample of phase unknown genotypes in simulation study.

featuring a family of size 1 correspond to unrelated individuals, or, if parental genotypes are provided, to parent-child trios[2]. One consequence of using this simulation strategy is that increasing the size of a sibship will reduce the number of independent haplotypes in a given sample. This allows for an evaluation of whether the resolving power from additional pedigree data compensates for the loss in independent haplotypes (i.e. whether the improved quality of the data compensates for the reduced quantity).

5.2.3 Summarizing the Results

Results from the simulation study are described using two standard summary statistics based on haplotype frequency estimates and phase accuracy. The Discrepancy metric (Excoffier and Slatkin, 1995; Kirk and Cardon, 2002) is used to assess haplotype frequency estimates. This is defined as:

$$D(\Theta;\hat{\Theta}) = \frac{1}{2}\sum_{i=1}^{2^L} \left|\theta_i - \hat{\theta}_i\right|. \tag{5.2}$$

For the phase configurations, the most likely phase configuration for each individual is calculated using the estimated haplotype frequencies. The percentage of individuals that are incorrectly assigned is then calculated. This metric is appropriate since it is the typical criterion on which haplotypes are assigned for use in a fine-scale mapping analysis. Results for these two measures of accuracy are presented in Table 5.3. The table is structured

[2]This is an attractive feature of the simulation strategy since studies based on parent-child trios are the most frequently encountered for all species.

to highlight a wide variety of trends, some of which are indexed by letters that will be referenced in the text. When referring to an entry in the table indexed by X, the notation **(X)** is used. Only the indices for the $APOE_1$ results are shown since annotating each table would have obscured trends. It will be contextually clear which distribution(s) are relevant to supporting a given statement. Similarly, standard errors are not included. Comparative statements were verified at the 95% significance level using a paired t-test.

The table also features results from a standard analysis using the EM algorithm for unrelated individuals (shaded column). These will be useful when discussing the results from treating related individuals as unrelated. As stated in Chapter 4, the model gives the same results as the EM algorithm for unrelated individuals when no family data is provided **(A)**. Hence either entry can be used to describe accuracy for 100 unrelated individuals, which is often useful as a base comparison to other scenarios that use family data.

Broadly, this study is concerned with how changes in family data, sample size and frequency distribution impact each of the two measures of accuracy. Since the number of progeny is fixed at 100, sample size is measured by the number of independent haplotypes segregating in the sample.

Section 5.2.4 focuses on the accuracy of haplotype frequency estimation. Section 5.2.5 provides similar analysis for phase reconstruction accuracy, while also highlighting how the two metrics differ in sensitivity to family data. These two sections collectively illustrate the importance of the true haplotype frequency distribution in determining the magnitude of reconstruction error as well as the effectiveness of reallocating resources for family data.

5.2.4 Impact of Family Data on Haplotype Frequency Estimation

First, note that for a given family size, increasing family information typically results in an improvement in accuracy (i.e. discrepancy decreases along a given row). However, adding family information does not always contribute to accuracy, as can been seen in the case of adding the genotype from a single parent **(C)**. This is because the number of progeny is sufficient to explain the parental phase and therefore the sire genotype provides redundant

HAPLOTYPE FREQUENCY DISCREPANCIES

	Family Size	UN	P	PS	PSD
$APOE_1$	1	0.116 **A**	0.116 **A,B,F**	0.078 **B**	0.063 **B**
	2	0.123	0.117	0.095	0.070
	5	0.143	0.119	0.109 **F**	0.077
	10	0.162	0.119 **E**	0.116 **E**	0.081
	25	0.212	0.122 **C**	0.122 **C,D**	0.082 **D**
$APOE_2$	1	0.082	0.082	0.061	0.049
	2	0.091	0.088	0.077	0.060
	5	0.110	0.097	0.090	0.065
	10	0.132	0.103	0.099	0.066
	25	0.180	0.104	0.104	0.070
$IL8_E$	1	0.047	0.047	0.037	0.032
	2	0.054	0.054	0.051	0.040
	5	0.065	0.060	0.058	0.043
	10	0.085	0.065	0.063	0.046
	25	0.110	0.068	0.068	0.047

PHASE RECONSTRUCTION ERROR RATE

	Family Size	UN	P	PS	PSD
$APOE_1$	1	0.21 **A'**	0.21 **B',F'**	0.12 **B',H**	0.06 **B',I**
	2	0.20	0.17	0.10	0.03
	5	0.19	0.12 **H**	0.06 **F',I**	0.02
	10	0.18 **J**	0.07 **J**	0.05	0.02
	25	0.15 **J**	0.06 **J**	0.06	0.02
$APOE_2$	1	0.13	0.13	0.08	0.04
	2	0.13	0.11	0.07	0.03
	5	0.13	0.09	0.05	0.02
	10	0.11	0.06	0.05	0.02
	25	0.11	0.05	0.04	0.02
$IL8_E$	1	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00
	25	0.00	0.00	0.00	0.00

Table 5.3: Impact of family size, family information and parental haplotype distribution on estimated haplotype frequencies, as measured by the discrepancy statistic (Top) and on phase reconstruction error (Bottom). Each entry in the table corresponds to the average value for 100 replicates. Shaded areas denote estimates that were obtained using the standard EM algorithm for unrelated individuals, boxed areas denote estimates obtained by treating related individuals as unrelated. Letters are referenced in the text.

information. By contrast, there is always an improvement in discrepancy if both parental genotypes are included **(D)**. This is because the second parental genotype will always provide information regarding an additional independent haplotype (which follows from our assumption of one progeny per dam).

This example demonstrates how increasing the number of independent haplotypes *or* the amount of family information improves accuracy. The question of whether resources intended for population data should be reallocated for family data is concerned with whether one should be increased at the expense of the other. This question was addressed for the case of nuclear families versus unrelated individuals in several studies, which showed that the optimal allocation decision will be frequency-dependent (Becker and Knapp, 2002; Schaid, 2002). These results illustrate that these frequency-dependent trade-offs between the quality and quantity of population data can be found for many pedigree configurations. Specifically, it is instructive to compare the accuracy of 200 independent haplotypes from a sample of unrelated individuals to 140 independent haplotypes segregating in 20 half-sib pedigrees of size 5 with a typed sire **(F)**. For the $APOE_1$ distribution, better accuracy is achieved from using more family data and fewer independent haplotypes, while for the $APOE_2$ and $IL8_E$ distributions more independent data is preferable to family data. As discussed in the previous section, a random sample generated from the $APOE_1$ distribution will have the most uncertainty associated with phase assignments and therefore will benefit most from family data.

Another important observation is that when family data is ignored (i.e. related individuals are treated as unrelated) discrepancy increases with family size **(G)**. This follows since increasing family size (i.e. increasing the number of conditional dependencies in the data) implies further deviation from the assumption of unrelated (independent) individuals. (This trend will be explored further in Section 5.3).

5.2.5 Impact of Family Data on Phase Reconstruction Accuracy

The most striking result is the uniformly perfect phase reconstruction given by $IL8_E$, which provides an example of a distribution where family data is redundant despite over 50% of a

sample containing ambiguous genotypes. These results are consistent with the frequency-known error rate given in Table 5.1. Indeed, for all three distributions, the observed error rate **(A')** is fairly close to the frequency-known error rate, which is the best-case average error rate that can be achieved when using the EM algorithm. In this context, it is reasonable to claim that EM-based phase assignments are accurate.

For the $APOE_1$ and $APOE_2$ distributions, increasing sib size and adding parental marker data always improves phase reconstruction accuracy. Specifically, as one move down a given column or across a row for either distribution, one observes a *gradual* decrease in phase reconstruction error rate. It should be noted that when resources are fixed, increasing family size decreases the number of independent haplotypes used in the subsequent study, and therefore this gain in phase reconstruction accuracy may not be justified.

While phase reconstruction error decreases with family information, it is not eliminated. Even for very large sib sizes, there is a small, but significant error when both parental genotypes are provided. Note also that for both distributions, there is also a discernible increase in the error rate when only the genotype for the common parent is provided. However, results for the $APOE_1$ distribution are consistently worse than for the $APOE_2$ distribution. These observations highlight the importance that both the frequency distribution and the pedigree structure have in determining whether resources should be allocated to ascertain family data.

Although increasing sib-size and parental marker information will both improve phase reconstruction accuracy, obtaining parental marker data is more efficient than adding more half-sibs. For both distributions, introducing genotype data for an untyped parent is more efficient than introducing as many as five additional half-sibs (**H,I**). This complements the results of Schaid (2002), which demonstrated that two full-sibs with untyped parents can be very inefficient in the context of optimal frequency estimation.

It is important to recognize that reconstruction accuracy for progeny does not extend to parents. This means that parental phase may still be incorrectly reconstructed even when reconstruction is accurate for progeny. Adding half-sibs will help reconstruct phase for an untyped common parent, yet our results show that the total number of half-sibs needed to make this parental genotype redundant can be quite large. For each of the three distribu-

tions, it can be seen that a typed sire still provides a small, but significant, improvement in discrepancy when as many as 10 progeny are available **(E)**. Introducing sibs without genotyping parents can actually be worse than reconstruction from population data if this information is to be used in a subsequent LD model that relies on haplotype accuracy of both parents and progeny (Meuwissen et al., 2002; Lee and van der Werf, 2004).

Note that the frequency-dependent trade-off between independent haplotypes and family size that was observed for optimal haplotype frequency estimation **(F)** is not applicable to optimal phase reconstruction accuracy **(F')**. Although increasing family information tends to improve both haplotype reconstruction accuracy and phase reconstruction accuracy the impact of family data differs. For example, the marginal gain in haplotype frequency accuracy from introducing the genotype of each parent decreases **(B)**, yet each parent makes roughly equal contribution to accuracy in the context of phase reconstruction accuracy **(B')**. There is an actual contradiction in trends between frequency estimation and phase assignment when related individuals are treated as unrelated. Specifically, it can be seen that phase accuracy *improves* as the number of related individuals that are treated as unrelated increases **(G')**. This paradox can be explained by noting that for more closely related individuals, there is a higher probability that two haplotypes are IBD (i.e. more homozygosity) in the genotype data. Fallin and Shorck (2000) made a similar observation when investigating the robustness of the EM algorithm to departures from Hardy-Weinberg Equilibrium by imposing homozygosity on the haplotype data. When individuals are related, however, the appropriate comparison is not the "benefit" in phase reconstruction accuracy relative to unrelated individuals, but the loss by not properly accounting for the conditional dependencies in the data. This loss can be quite large as seen in the case of $APOE_1$ where over 10 haplotypes are incorrectly assigned by not accounting for underlying pedigree structure **(J)**.

5.3 Simulation Study: Sampling from Small Hierarchical Populations

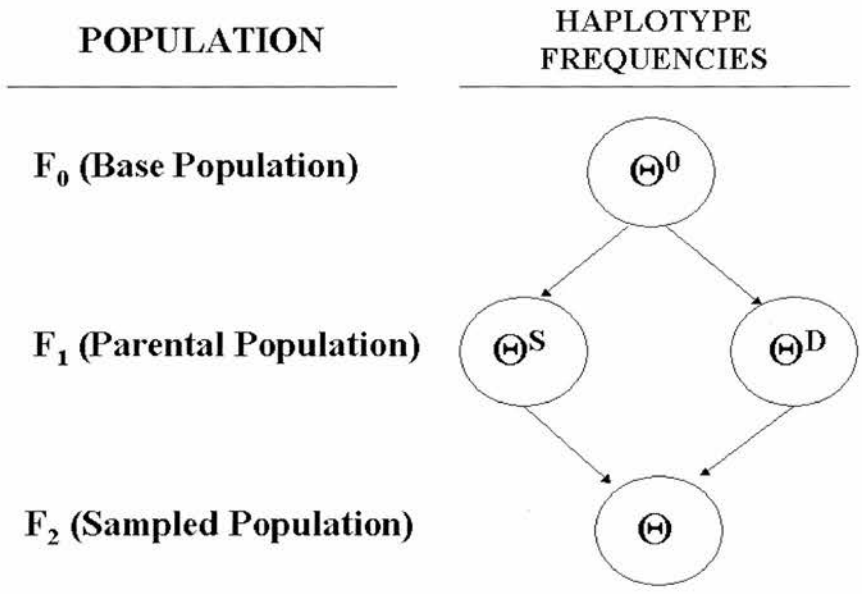
The previous section examined scenarios where a sample contained a specified number of sibships, each having equal size. This sample configuration is unrealistic if the sample is drawn from an outbred population. However, as is explained below, samples featuring an expected number of sibships are known may be encountered. This section examines the consequences of treating related individuals as unrelated on haplotype frequency estimation *when the number of sibships in a given sample is variable*. Before explaining the simulation strategy, the practical relevance of this scenario is discussed.

5.3.1 Overview

Most statistical models that reconstruct haplotypes from population data are based on assumptions that are applicable only for large, panmictic populations. These assumptions include the same haplotypic structure for both sexes, and that a random population sample will feature only unrelated individuals. When the target population is derived from a smaller, hierarchical population, a random population sample may feature individuals who are related through a common parent. In practice members of the sample are treated as unrelated, even when there is a high probability that the kinship coefficient between any two individuals is high. It is therefore important to investigate whether ignoring relatedness in this context adversely impacts the accuracy of haplotype frequency estimates.

To conduct this simulation study, it is necessary to simulate data that conforms to an appropriate population structure. Section 5.1 describes a hierarchical population that is often encountered in livestock genetics and develops the appropriate model to accommodate the structure. The simulation strategy is discussed in Section 5.3.3 and results are presented in Section 5.3.4.

Figure 5.1: Conceptual Framework for Evaluating Population Structure and Relevant Population Parameters.



5.3.2 The Sampled Population

Figure 5.1 depicts a standard breeding scheme that is commonly employed for many species of livestock. Formally, the target population (which is denoted F_2) is derived from the the random union of gametes from N_S sires and N_D dams, which collectively comprise the parental population, denoted F_1. Members of F_1 are selected from a base population, denoted F_0, that is assumed to be in Hardy-Weinberg Equilibrium (HWE). An additional assumption is that the markers under consideration are unlinked with any genes that may influence selection criteria. This last assumption allows us to describe the difference between F_0 and F_2 exclusively in terms of N_S and N_D. Specifically, when both N_S and N_D are large, F_2 approximates the Hardy-Weinberg model that defines F_0.

N_D is assumed sufficiently large so that N_S can be regarded as the single measure by which F_2 violates HWE. Under this additional constraint, members of a random population sam-

ple are either paternal half-sibs or unrelated.

Figure 5.1 also illustrates that there are now four parameters that will be relevant for inference. Recall that in a model which assumes HWE, the sole parameter of interest is $\Theta = (\theta_1, \ldots, \theta_M)$. The additional parameters that are relevant to small hierarchal populations derived from this breeding design are Θ^S, Θ^D and Θ^0, which denote haplotype frequency vectors in the N_S sires, N_D dams and base population respectively[3].

Under the assumption of a randomly mating parental generation, all relevant haplotypic information for F_2 can be summarized in terms of Θ^S and Θ^D. If the primary interest is in actual phase reconstruction, the posterior probabilities, $p(z|y, \Theta^D, \Theta^S)$, follow from:

$$p(z = (h_j, h_k)|\Theta^D, \Theta^S) = \begin{cases} \theta_j^D \theta_k^S + \theta_j^S \theta_k^D & j \neq k, \\ \theta_j^D \theta_k^S & j = k \end{cases} \tag{5.3}$$

while if the interest is in the haplotype frequencies segregating in F_2, one would use:

$$\Theta = \frac{\Theta^S + \Theta^D}{2}. \tag{5.4}$$

Hence, the objective of inference is to estimate Θ^S and Θ^D.

The critical assumption made in this breeding design is that the parental haplotypes are conditionally independent given Θ^0. The complete-data log likelihood is analogous to the likelihood for the model presented in Chapter 4:

$$\begin{aligned} \log L_C(\Theta^0) &= \log\left\{p(\mathbf{s}, \mathbf{d}, \mathbf{y}|\Theta^0)\right\} \\ &= \sum_{i=1}^{P} \log p(s^i|\Theta^0) + \sum_{j=1}^{N} \log p(d_{j.}|\Theta^0) + \sum_{k=1}^{N} \log p(y_{ki}|s^i, d_{ki}), \end{aligned} \tag{5.5}$$

Inference for Θ^0 can therefore be conducted efficiently as described in Chapter 4.

Once $\hat{\Theta}^0$ is obtained, it can be used to estimate Θ^D under the assumption that N_D is large. Deriving $\hat{\Theta}_S$ requires an additional step. Let $\alpha = \frac{P}{N_S}$ be the proportion of sires observed. $\hat{\Theta}_S$ can be calculated as a weighted average of the observed sires and the unobserved sires,

[3]The reason why these additional parameters are not required under HWE, is that the haplotype frequency profile is constant in successive generations, i.e. $\Theta^0 = \Theta^S = \Theta^D = \Theta$.

estimated by $\hat{\Theta^0}$:

$$\hat{\theta}_j^D = \hat{\Theta}_j^0 \qquad \hat{\theta}_j^S = \alpha \sum_{i=1}^{P} E_{p(s^i|\mathbf{y}_{.i},\hat{\Theta^0})}\left[n_{ij}\right] + (1-\alpha)\hat{\theta}_j^0 \tag{5.6}$$

for $j = 1 \ldots M$. Once $\hat{\Theta}^S$ and $\hat{\Theta}^D$ are obtained, they can be used to estimate $\hat{\Theta}$ according to equation (5.4).

5.3.3 Simulation Strategy

All haplotype frequencies used in this simulation study are based on the study by Hull et al. (2001). This includes the haplotype frequency estimates for the European sample that was used in the previous study ($IL8_E$). Haplotype frequency estimates for an African population are also employed. As reported by the study, only 12 of the possible 64 haplotypes are segregating at these loci. Additionally, a hypothetical population is considered where the 12 observed haplotypes occur with equal frequency. This gene diversity statistics for the three respective frequencies are .5619 (European), .7823 (African) and .9167 (Hypothetical).

In this simulation study, the number of sires is held fixed and the sibsize is allowed to vary. Specifically, to simulate each member of the sample, a sire is selected at random from a fixed set of sires and then mated to a dam (which is generated from two draws of the haplotype distribution). Haplotype reconstruction is then conducted when the pedigree structure is known and when the pedigree structure is unknown, and members of the sample are assumed unrelated. In all scenarios, parents are assumed untyped.

As stated above, there are four different frequency parameters that are relevant in this simulation study. The results for Θ^0 (the base population) and Θ (the target population) are presented[4].

A wide variety of sire/sample size configurations were explored. For each configuration, a test was conducted to determine whether the difference between the discrepancy obtained

[4]Recall that Θ is a function of the sire and dam frequencies, and therefore estimates for these two parameters are omitted.

by UN-EM and the discrepancy obtained from incorporating the sparse pedigree information is significantly greater than zero. It is meaningful to report the percentage decrease in discrepancy attained from using the correct model specification over the standard EM-based approach where all individuals are assumed unrelated (UN-EM). The results, which are presented in Table 5.4, can therefore be interpreted as the "gain" from using the more accurate model for a given configuration. Equivalently, these results capture the robustness of a statistical model that assumes independence to the dependencies in a random sample.

5.3.4 Results

Table 5.4 highlights three important trends. First, for a given diversity and sample size, the gain from using the correct specification decreases with the number of sires. This follows since the probability of selecting two related individuals at random decreases as the number of parents increase. As stated above, if there is no relatedness in the sample, the model generates the same results as UN-EM. Qualitatively, the number of sires is a measure of how the underlying population structure deviates from HWE, i.e. the more sires contributing in the parental generation, the closer the population structure approximates HWE. This is illustrated by the converging lines in Figure 5.2 which provides a graphical perspective for this trend.

The second trend that is captured by Table 5.4 is that, for a given number of sires and diversity, the percentage decrease in discrepancy becomes more pronounced as the sample size increases. This can also be seen in Figure 5.2, where the discrepancy gap between the two different model specifications is wider for the larger sample size.

The third trend is that discrepancy becomes more pronounced as gene diversity increases. This can be attributed to (a) higher sampling variance that arises from the more uniform haplotype frequency distribution suggested by the high diversity statistic and similarly (b) more ambiguous genotypes, which are likely to benefit from any pedigree information.

High Gene Diversity

Sample Size \ Sires	25	50	75	100	150
20	0	1	-2	-1	0
50	7***	3	0	0	1
100	14***	6***	5**	5**	1
150	17***	11***	6***	3*	4*

Sample Size \ Sires	25	50	75	100	150
20	0	1	-2	0	0
50	7***	3	0	1	1
100	12***	6***	4**	5**	1
150	18***	9***	6***	5**	3

Moderate Gene Diversity

Sample Size \ Sires	25	50	75	100	150
20	0	3	2	0	-1
50	6**	4	1	3	3
100	12***	6**	5*	4	1
150	16***	8***	7**	6**	1

Sample Size \ Sires	25	50	75	100	150
20	-1	3	3	-1	-1
50	5	5*	2	3	3
100	13***	6**	5*	3	0
150	15***	7**	5*	5**	2

Low Gene Diversity

Sample Size \ Sires	25	50	75	100	150
20	2	3	-1	1	0
50	0	0	0	1	0
100	5	-2	-1	2	0
150	12***	6	7*	3	1

Sample Size \ Sires	25	50	75	100	150
20	2	3	-1	1	0
50	3	0	1	1	0
100	6*	1	1	2	1
150	13***	5	7*	3	2

Table 5.4: Percent reduction in discrepancy for Θ^0 (left) and Θ (right) attributable to modelling dependencies in random population sample. * denotes significance at the 10% level, ** denotes significance at the 5% level and *** denotes significance at the 1% level.

5.4 Discussion

The results of this study have significant implications for an experimental design using two-stage haplotype analysis[5]. The effectiveness of a two-stage haplotype analysis will be contingent on two factors: 1) the magnitude of the estimation error and 2) the sensitivity of the subsequent haplotype-based analysis to this estimation error, which can be determined from simulation studies. This study has shown the magnitude of the estimation error depends on the the haplotype frequency distribution[6].

[5]Two-stage haplotype analysis first entails inferring haplotypes from a sample of phase-unknown genotypes using a computational algorithm. The second stage entails using these haplotypes in a multi-locus LD model, where they are treated as having been directly observed.

[6]The second factor is addressed in Chapter 6.

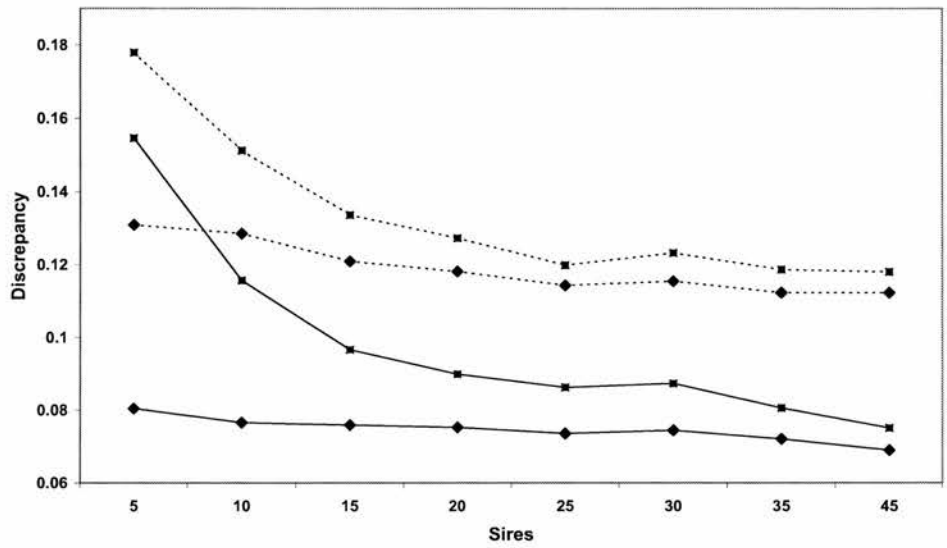


Figure 5.2: Impact of Sample Size, Number of Sires and Model Specification on Discrepancy of Θ. Results generated from simulations on haplotypes exhibiting moderate diversity (results will be more/less pronounced for higher/lower gene diversities). Squares denote mean discrepancy from 250 simulations using UN-EM while diamonds indicate similar results using our model specification. Dashed lines correspond to a sample size of 50 and solid lines correspond to a sample size of 150.

It was also shown that while reconstruction errors may be unavoidable (i.e. independent of sample size), this error rate can be calculated directly from the frequency distribution. An example is provided where the most likely phase criterion will yield perfectly accurate results, even for a sample containing a large proportion of ambiguous genotypes. Practically, it is possible to calculate the error rate from an estimated frequency distribution that is based on a preliminary sampling of the population. This estimated error rate can then be compared against a predetermined threshold denoting the minimum level of phase accuracy. If the predicted phase reconstruction error exceeds this threshold, then either pedigree data is required or an alternative to the two-stage approach must be used. The effectiveness

of pedigree data will depend on the haplotype frequency distribution, as well as the type of pedigree information provided. Two kinds of information are considered: increasing the family size and introducing parental marker data. For half-sibs, introducing parental genotypes is more efficient than increasing family size and leaving parents untyped. In general, assessing the effectiveness of pedigree structure and frequency distribution on haplotype inference can be done through a simulation study based on the inferred frequencies.

EM-based haplotype frequency estimates are often considered accurate for sample sizes consisting of approximately 100 individuals (Fallin and Shorck, 2000; Qin et al., 2002). For smaller samples sizes, or for samples featuring large amounts of missing genotype data, a Bayesian approach with an informed prior, such as the coalescent prior introduced by Stephens et al. (2001), may be appropriate. However, as shown in Chapter 2, the EM algorithm appears to be reliable even for very small sample sizes. The EM algorithm is also considered to provide accurate results under departures from the random-mating assumption (Fallin and Shorck, 2000). However, our results demonstrate that failing to account for related individuals in a sample (by treating each member in the sample as independent) can lead to an appreciable loss of efficiency. Our results demonstrate that accounting for small sibships from untyped parents in a random sample can result in a significant improvement in accuracy of population parameters, including haplotype frequency estimates for the target population. Importantly, it is shown that even samples that feature sparse relatedness (i.e. small sibships with untyped parents) can yield significantly better haplotype frequency estimates using the model proposed in the previous chapter than UN-EM, which assumes no relatedness in the sample. Hence, it is important to ascertain pedigree information, e.g. by sibship reconstruction methods described by Thomas and Hill (2002) in the context of haplotype analysis.

Chapter 6

Haplotype Analysis in Association Studies

This chapter provides an overview of the utility of haplotypes in fine-scale mapping. As noted earlier, a primary reason for the interest in haplotypes is because they may provide critical information for identifying complex disease variants or quantitative trait loci (QTL) in an LD mapping procedure. Haplotypes play two important roles in LD mapping. First, they can help establish whether there is a sufficient relationship between physical distance and LD in the target population. This is a necessary criterion for any LD mapping approach to be effective. Second, it is believed that haplotypes can optimize the power and accuracy in the actual identification of complex trait variants (assuming necessary criteria are met so that the variants can, in principle, be found).

This chapter is divided into two parts. The first part provides an empirical assessment of association studies. It lists possible reasons for the failure of so many association studies to detect disease variants. It also provides a simulation study that may help explain a reason for the high level of false positives in published studies. The second part of the chapter provides a brief evaluation of how haplotypes may improve the effectiveness of an association analysis.

6.1 An Empirical Assessment of Association Studies

Association studies are the standard approach for fine-scale LD-mapping. To identify a causative variant for complex disease, an association study simply tests whether certain alleles occur more frequently in a sample of unrelated affected individuals than in a sample of controls. Statistically significant differences are interpreted to indicate the presence of a nearby disease variant.

Association studies were initially regarded as an enormously appealing paradigm for mapping complex disease variants. This is because they are simple to implement, cost effective and, most importantly, they are generic (i.e. they do not require any knowledge of the biology of disease). In practice, however, the performance of association studies has been disappointing. First and foremost, the vast majority of association tests have failed to identify any disease variants. Indeed, while thousands of association studies have been conducted for many complex diseases, only 50 genes and their allelic variants can be considered true positives[1](Wang et al., 2005). Another disappointment is that the majority of variants that were identified and published turned out to be wrong, i.e. false positives. It is estimated that 75% of published associations failed to be replicated in subsequent studies (Lohmueller et al., 2003)[2].

The next two sections discuss some of the reasons that may contribute to these alarming statistics.

6.1.1 Failure to Detect Disease Variants

In order for an association study to be successful, three criteria must be met: The disease must have a genetic basis; Common mutations underlie common diseases; and, there must be appropriate patterns of LD. This section begins by discussing each of the criterion.

[1]This should be contrasted with the 1200 confirmed variants that cause simple, monogenic disease (Botstein and Risch, 2003).

[2]Note that associations here refer to associations of a genomic region with a complex trait and not necessarily associations of a particular variant within that region. Even if an association with a particular region has been established, there is still several more stages in fine-mapping before a particular variant is ascertained. This would explain the fact that only 50 true variants have been validated, while quite a few more genomic regions have been implicated through replicated studies.

There is considerable debate over whether the second and third criteria hold. It concludes with an assessment of what can be/what is being done to resolve debate.

6.1.1.1 Three Fundamental Criteria for LD Mapping

1. The disease must have a genetic basis

The first criterion is that there is, indeed, a genetic basis for the target disease (and that it is not, e.g., caused by an infectious agent). Specifically, there must be at least one polymorphic locus where an allele confers a higher relative risk for the disease. Family studies are typically used to establish whether a disease has a genetic basis. This is because relatives share a greater proportion of their genes than unrelated individuals, and therefore genetic diseases tend to cluster in families.

There are many well established tests to establish that a disease has a genetic basis. The failure of association studies cannot be attributed to incorrectly assigning a genetic basis to a disease when there is, in fact, none.

2. Common mutations with sufficiently large effect are behind common diseases

As noted above, association studies detect whether there is an excess of allele sharing among cases. Because the disease is complex, not all cases will exhibit the disease phenotype for the same reasons. It is clear that for an association study to work, at least some of the cases must share the same disease variant. Furthermore, the disease variant must have sufficiently strong effect so that it can be detected through the "statistical fog" created by the other risk factors that, by definition, contribute to complex disease. The standard measure of allelic effect is the odds ratio (OR), which is defined as the odds of exposure to the genetic variant in cases compared with that of controls.

Ideally, then, variants will confer high OR *and* be common in the population. The rarer the variant and/or the lower the OR, the weaker the linkage signal and the larger the sample size[3] needed to indicate significance[4]. Most case-control studies ascertain fewer than

[3]An OR of less than 1.2 are considered difficult to reliably detect with realistic sample sizes(Zondervan and Cardon, 2003).

[4]Therefore, these tests do have the power to detect rarer variants, provided the OR is sufficiently large. From a medical standpoint, it is preferable for the variants to be common (i.e. the results would impact a

1000 cases and controls, reflecting the assumption that allelic effects of variants underlying complex disease have an OR larger than 2 (Zondervan and Cardon, 2003) and a population frequency larger than 30% (Weiss and Clark, 2002).

The critical assumption that common mutations are behind common diseases is known as the Common Disease Common Variant (CDCV) hypothesis. Implicitly, these mutations must also confer sufficiently large effect. There is a growing debate over whether this hypothesis is valid. Critics argue that the average OR of the mere 50 confirmed disease variants that have been mapped is less than 2. In fact, until recently the only variant that conforms to the CDCV hypothesis is the APOE variant implicated in Alzheimer's Disease, with a population frequency of 15% and an allelic OR of 3.3.

There was much excitement when a variant for Age Related Macular Degeneration (AMD) was recently mapped using a standard case control design (Klein et al., 2005). The variant had an OR of around 3 and a population frequency of 40%. Paradoxically, the success of mapping the AMD variant may weaken the CDCV hypothesis. This is because the statistical methods used to detect the AMD variant are similar to those employed in the thousands of studies that failed to detect variants for other complex disease. Equivalently, if most common diseases had similar profiles, there should have been more confirmed associations than the 50 cited above.

3. Case chromosomes must exhibit LD in the region surrounding the disease variant(s)

A key assumption for association studies to work is that a sample of case chromosomes should exhibit an excess of allele sharing for markers surrounding the disease variant. This implies that markers will exhibit LD in this region. This assumption is necessary for an effective association study since the actual disease variant is not directly tested. Specifically, only a small subset of alleles within the target region are actually used[5]. These are chosen on the basis of logistical considerations rather than on any prior belief of function, and hence are not assumed to contain the disease variant.

larger proportion of the population). It should also be noted that rare variants with very large effects can be considered Mendelian traits, and have been mapped using with conventional (pedigree-based) linkage analysis. These variants are considered "low hanging fruit".

[5]It would not be economically feasible to test all alleles.

The reason that this assumption may be valid is because, *provided the CDCV hypothesis holds*, most affected individuals will have inherited the causative allele from the same founder. Because crossover points are sparsely distributed during each meiotic event, the alleles surrounding the variant are also likely to be IBD with the founder alleles. In the case of LD-based mapping, the number of meiotic events between the founder (a distant ancestor) and cases will be considerably larger than in linkage mapping (where the founder, a parent, is only one generation removed from the affected cases). Hence, the IBD haplotype will span a much shorter distance than in linkage, allowing for fine-scale mapping.
Skeptics argue that the relationship between LD and physical distance is likely to be confounded by a myriad of other factors as the population evolves. However, recent empirical evidence has established the presence of well defined haplotype blocks. Specifically, a variety of studies have indicated that 70-80% of the genome has regions of high LD (Wang et al., 2005), and that these regions can be divided into blocks spanning an average 200kb that exhibit limited haplotype diversity. Since the LD within each block high, it should be necessary to select only a subset of "tag" SNPs from each block, which could act as a surrogate for other SNPs.

6.1.1.2 Resolving the Debate

Genome wide associations test are based on uniformly distributed markers are not efficient since the patterns of LD vary widely throughout the genome. To establish the existence of the CDCV hypothesis and appropriate LD, there are three alternative strategies that could be useful: first, candidate genes, or regions, could be targeted and sequenced in their entirety; second, association studies could utilize the entire set of SNPs in the genome; and third, patterns of LD for major populations could be established empirically, and then a subset of tag SNPs could be ascertained. The first two approaches would be a concerned with determining whether the CDCV hypothesis is valid (the nature of LD is irrelevant since the causative variant itself would be tested). However, neither strategy is likely to be realistic in the near future: The candidate gene approach requires prior knowledge of the biochemical basis for the trait, and such knowledge is notoriously elusive. A comprehen-

sive genome-wide scan of the more than 3 million SNPs is simply not economically viable, even with rapid advances expected in genotyping technology.

This leaves the final option, which is currently being undertaken by the HAPMAP consortium. Identifying all haplotype blocks and tag SNPs is the central objective of the HAPMAP project (Weiss and Clark, 2002). Proponents of the HAPMAP project argue that failure of association studies is not necessarily because the CDCV hypothesis doesn't hold, but because SNP selection was currently inefficient. Completion of the HAPMAP project in 2007 will ideally provide a set of tag SNPs that will cover (i.e. be in high LD with) all other SNPs in the genome. If the project is successful, then the CDCV hypothesis can be subsequently resolved.

6.1.2 The High False-Positive Rate

Most association studies fail to detect any variant for reasons stated in the previous section. However, of the variants that have been reported, 75% turned out to be false-positives, i.e. results of these studies were unable to be replicated in other studies. Two of the most popular reasons given for the high false-positive rate are failing to account for multiple testing and failing to account for population stratification.

The above statistics refer to human populations. In animal populations, samples that are treated as unrelated may actually contain related individuals. This section examines the impact of treating related individuals as unrelated on the false-positive rate, or type I error, of a standard association test. Specifically, it examines the scenario where a case-control study is carried out and the members in the sample are not all unrelated, as required by the test. (Recall that a standard case-control association test determines whether differences between either marker or haplotype frequencies of randomly selected cases and controls are significant). It should be noted that this scenario is relevant for many natural and domestic populations, where, in the absence of pedigree information, a sample may be treated as unrelated even through the sampled members can be closely related.

Treating related individuals as unrelated might be expected to increase the probability of committing a Type I error, since variability that naturally arises between two groups sam-

	Family Size	Single Locus Analysis for Hardy-Weinberg Disequilibrium[1]		Association Test using Haplotype Frequency Estimates	
		$\chi^2(1)$	**sdev**	χ^2 **(8)**	**sdev**
$IL8_E$	**1**	1.69	2.23	9.32	4.05
	5	1.71	1.87	14.53 *	6.57
	10	1.67	1.74	18.78 **	9.22
	25	2.17	2.42	26.17 ***	16.37
	50	3.55 *	4.27	--	--
		$\chi^2(1)$	**sdev**	χ^2 **(9)**	**sdev**
$APOE_2$	**1**	2.47	2.05	9.37	4.12
	5	2.29	1.99	14.76 *	7.09
	10	2.24	1.87	21.35 **	9.69
	25	2.79	2.61	35.73 ***	17.91
	50	4.50 **	4.42	--	--
		$\chi^2(1)$	**sdev**	χ^2 **(12)**	**sdev**
$APOE_1$	**1**	2.43	2.11	12.51	4.66
	5	2.43	1.99	21.41 *	7.47
	10	2.36	2.00	27.66 ***	12.24
	25	2.99	2.77	47.68 ***	19.77
	50	5.58 **	4.95	--	--

[1] For each replicate, the test statistic for the locus exhibiting the highest disequilbirum was used.

Table 6.1: Impact of treating half-sibs as unrelated on two nonparametric tests: single-locus tests for Hardy-Weinberg Disequilibrium (left) and non-homogeneity of haplotype frequency profiles for case control data in a neutral genomic region (right).

pled from the same neutral region will be accentuated when the data are dependent. To investigate this, we simulate neutral marker data for hypothetical cases and controls according to the procedure outlined in Sections 5.2.1 and 5.2.2. Briefly, we are considering a sample featuring a collection of paternal half sibships. The total number of sampled individuals is fixed at 100 and sib-sizes are fixed at 1,2,5,10 and 25. Parental phase configurations are determined by independent draws from the haplotype frequency distributions given in Table 5.1. The frequency distributions should be distinguished according to their

gene diversity, with $APOE_1$ having the largest diversity and $IL8_E$ having the smallest diversity.

In the simulation study, frequencies are estimated under the assumption that individuals are unrelated using the EM algorithm. The chi-square statistic is then calculated based on the reconstructed frequencies. The results are presented in the right column of Table 6.1. They reveal that an inflated Type I error is likely to be realized for a family size of 5 or larger for each of the distributions. Even though the actual discrepancy statistic is uniformly greatest for the $APOE_1$ and least for $IL8_E$, the rate at which significance increases with family size is fairly consistent for each of the three distributions: the principal difference lies in the standard deviation. We also ran a standard single locus Hardy-Weinberg test for each of the loci and present the results on the left of Table 6.1. This reveals that data sets that are sufficiently related to exhibit a Type I error in case-control association analysis will not be detected using the standard single locus Hardy-Weinberg test.

These results demonstrate that failing to account for related individuals in a sample (by treating each member in the sample as independent) can lead to an appreciable increase in Type I error in a study where two populations are contrasted using EM-based haplotype frequencies. The problem would have been avoided if dependencies in data were accounting for using the model that was developed in Chapter 4.

6.2 Using Haplotype Data in a Mapping Analysis

If the criteria listed in Section 6.1.1. are valid, then there are a number of valuable simulation studies that could be conducted to assess the optimal way to use haplotypes in a mapping analysis. These studies would help provide answers to the following questions:

1. Do haplotype-based tests provide more power than analogous single-locus procedures?

2. Do model-based approaches, which attempt to explicitly model the evolutionary history of the variant, provide more power than model-free approaches?

The following provides a brief review of the relevant research that has been conducted in these areas:

1. Single Markers or Haplotypes?

When considering whether haplotypes will provide more power over analogous single marker procedures, it is critical to distinguish whether a parametric or nonparametric approach will be employed. It is tempting to argue since haplotypes are more informative than genotypes, then using them must can only be beneficial, irrespective of the modeling approach. This would be incorrect if standard nonparametric procedures (e.g, the chi-squared test in Chapter 6) are employed. To understand this, consider an association test where, instead of individual markers, the haplotype spanning the entire chromosome is used. The haplotypes for all cases (and controls) are likely to be unique and the test will fail as there is no sharing between any affected individuals. This should be considered when attempting to reconcile the conclusion of Long and Langely (1999), which categorically stated that single marker tests are *more* powerful than haplotype-based tests, and a similar study by Grapes et al. (2004), that concluded that haplotypes were more powerful. The haplotypes used in the Long-Langely analysis spanned 20 markers, while the Grapes analysis used two locus haplotypes.

Many studies have been conducted to investigate the optimal haplotype size in nonparametric tests. In the standard chi-square test for association, the degrees of freedom increases with the number of haplotypes. It has been shown that large degrees of freedom can both inflate type I error (Fan and Knapp, 2003) and reduce power (Chapman et al., 2003). While, the optimal haplotype size will be contingent upon the pattern of LD in the relevant region, it has been shown that in regions of high LD, tests using haplotypes of 2-5 markers are more powerful than tests using single markers (see, e.g. Nielsen et al., 2004; Fan and Knapp, 2003; Zaykin et al., 2001).

In summary, the optimal haplotype length will be contingent on the age of the variant (and the extent of LD) which is impossible to estimate. Simulation studies should explore this further.

2. Model-Based or Model-Free Methods?

Recently a variety of model-based methods have been introduced as an alternative to non-

parametric tests. These attempt to effectively model of the mechanisms generating allele sharing around the variant.

In principle, model-based approaches can be expected to provide more power *provided that the model is a sufficient approximation of reality*. The majority of haplotype-based models (see, e.g. Lu et al., 2003; Morris et al., 2002; Perez-Enciso, 2003) use an HMM structure similar to the Lander-Green algorithm: each node corresponds to a marker, the observed node is the genotype and the latent node is an IBD state. However, with Lander-Green, the IBD state indicates inheritance from parents, and the transition probabilities are a function of the (known) recombination rate. With the LD-based models, the transition probabilities are a function of the time since the mutation was introduced as well as other parameters that are not known. This can adversely affect power. Comparisons with standard model-free alternatives are needed.

Chapter 7

Conclusions and Future Work

This chapter features a summary of the key contributions of this thesis, and provides some suggestions for future related work.

7.1 Summary

Contributions of the thesis include:

A New and Necessary Model for Haplotype Reconstruction
The thesis has provided a model for haplotype reconstruction that accounts for dependencies that may arise in a sample of genetic-marker data in current population studies. Such a sample can be characterized by two types of dependencies: dependencies that exist between markers (i.e. LD) and those that exist between individuals in the sample (i.e. family information). The thesis thus bridges the gap between the two standard classes of haplotype reconstruction models: those that account for LD but assume individuals are unrelated, and those that account for related individuals but assume markers are in Linkage Equilibrium. We expect the model will be useful beyond this thesis. This is because simulation studies based on empirical data have clearly underscored the importance of accounting for both types of dependencies when they are present. Equivalently, we have shown that existing haplotype reconstruction algorithms are not robust to violations of either assumption.

These analyses were principally concerned with the accuracy of haplotype frequencies and phase. However, we have also confirmed that the biased haplotype frequency estimates that result from treating related individuals as unrelated can impact association analysis, and in particular can inflate the false-positive rate.

A Powerful Computational and Conceptual Framework for Haplotype Analysis

The thesis demonstrated that real synergies exist between Machine Learning (and in particular the Graphical Model formalism) and Population Genetics. By employing the Junction Tree Algorithm, it was possible to efficiently conduct inference over important pedigree configurations that could not be accommodated by existing methods. Equally important was the insight gained by representing the likelihood in compact graphical form. By using graphical models as a visual, as well as computational, tool we were able to develop a truly unified model that bridged the ten-year gap between algorithms developed for unrelated individuals and algorithms developed for single pedigrees. This motivated a broad range of simulation studies that dispelled common misconceptions regarding the utility of family data in haplotype reconstruction (e.g. that a fixed number of sibs could always resolve phase).

Valuable Insight into the Role of Family Data in Haplotype Reconstruction

A commonly encountered question for many experimental designs is: How much family data (if any) should be used to facilitate haplotype reconstruction in a population study? We find that the impact of pedigree data depends upon the actual haplotype structure in the population. Since this structure will vary throughout the genome, there is unlikely to be a single optimal strategy for accurate haplotype reconstruction from markers used in genome-wide scans. Rather, the impact of various pedigree data on particular subsets of markers can be assessed through a simulation study based upon initial frequency estimates.

7.2 Future Work

Extensions to other pedigrees

The thesis focused on unrelated individuals, parent-child trios and arbitrarily large half-sibships. An obvious next step is to use the graphical model formalism to accommodate other pedigree structures. One of the key benefits of using graphical models is that they employ general purpose algorithms that apply to any graph. As was shown in Figure 4.5, the process of conducting inference on full-sibships is trivial once the half-sib model is understood. Furthermore, bounds on the computational complexity can be read straight from separator potentials.

An excellent opportunity for synergy between the two disciplines would be to generalize the model to accommodate inbred pedigrees. Inference over 'loopy" graphs is an active area of research in Machine Learning, and analogous pedigree structures are very important in many application areas in population genetics. Integrating this theory with the theory already developed for efficient calculation on complex pedigrees in statistical genetics (see,e.g. Lange and Elston, 1975; Cannings et al., 1978; Lange and Boehnke, 1983) could be very useful.

Genotyping Error and Recombination

Although the quality of marker data is improving, genotyping error is an unpleasant reality. When recombination occurs in a pedigree, this can be considered tantamount to a genotyping error if the model explicitly assumes no recombination.

Few studies have been devoted to investigating the impact of genotyping error on haplotype reconstruction, yet several empirical studies that were conducted during this thesis demonstrated how severe this impact may be. One one occasion, haplotypes for five biallelic loci were reconstructed using the standard EM algorithm. It was later discovered that one of the markers was mistyped. When the correct genotypes were used for this marker, the haplotype frequency profile for the other four markers changed dramatically.

Partition Ligation

The partition-ligation algorithm proposed by Niu et al. (2002) is becoming quite popular

for coping with large numbers of markers (see,e.g. Zhang et al., 2005; Qin et al., 2002). As noted in Chapter 2, more work needs to be done to establish the reliability if this procedure and the reliability of similar procedures mentioned in Chapter 2 that have been developed to accommodate large number of markers.

Appendix A

Optimal Experimental Design Revisited

In Chapter 5, it was shown that the tradeoff between family data and independent haplotypes was nontrivial when estimating haplotype frequencies. This appendix demonstrates that tradeoff can be trivial for similar graphical structures when the underlying distributions are different.

Consider the joint distribution given in Figure (A.1), with local conditional probability distributions:

$$p(x) \sim N(0, \sigma_x^2) \tag{A.1}$$

$$p(y_i|x) \sim N(x, \sigma_y^2) \qquad i = 1 \ldots n_S \tag{A.2}$$

$$p(z_{ij}|y_i) \sim N(y_i, \sigma_z^2) \qquad j = 1 \ldots n_p \tag{A.3}$$

The objective is to choose n_s and n_p that will provide the best estimate for x, subject to the constraint that $n_S n_p = N$. n_s is analogous to the number of sibships and n_p is analogous to the number of sibs per pedigree.

How should 'best' be defined? In Chapter 5, it was the combination of n_s and n_p that minimized the Discrepancy metric. Here, it will be the combination of n_s and n_p that minimize $\sigma^2_{x|\mathbf{z}}$, where

$$p(x|\mathbf{z}) \sim N(\mu_{x|\mathbf{z}}, \sigma^2_{x|\mathbf{z}}). \tag{A.4}$$

Hence, the objective is to specify $p(x|\mathbf{z})$ as a function of n_p and n_S and then evaluate the solutions, n_p^* and n_S^* that minimize the expression.

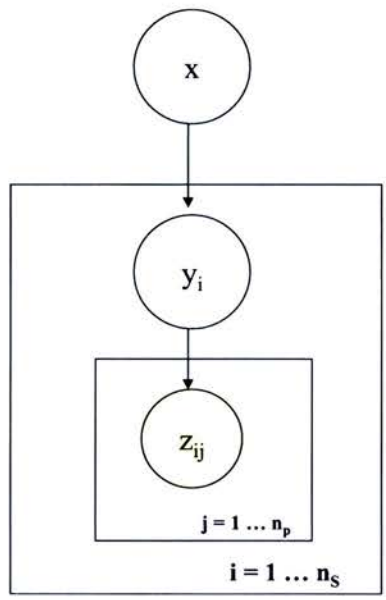


Figure A.1: Gaussian model to investigate trade-off between reduction in uncertainty (n_p) and sample size (n_S). x is the parameter of interest, y is the latent data, z is the observed data.

$p(x|\mathbf{z})$ can be derived by first specifying the joint distribution depicted in Figure (A.1):

$$\begin{aligned} p(x,\mathbf{y},\mathbf{z}) &= p(x)\prod_{i=1}^{n_S} p(y_i|x)\prod_{j=1}^{n_p} p(z_{ij}|y_i) \\ &\propto \exp\left\{\frac{x^2}{\sigma_x^2}+\frac{\sum_{i=1}^{n_S}(y_i-x)^2}{\sigma_y^2}+\sum_{i=1}^{n_S}\left(\frac{\sum_{j=1}^{n_p}(z_{ij}-y_i)^2}{\sigma_z^2}\right)\right\} \end{aligned} \tag{A.5}$$

From (A.5) it is possible to derive $p(x,\mathbf{z})$, which is Gaussian. Hence, the term in the exponent as a quadratic form in x and the mean and variance of (A.4) can be obtained by completing the square.

Obtaining $p(x,\mathbf{z})$ from (A.5) requires integrating over $\mathbf{y}$. This can be done efficiently by

using a peeling algorithm. For each "family", $i = 1 \ldots n_S$, remove (peel) those terms in (A.5) that characterize the following distribution:

$$p(y_i|x, \mathbf{z}, \mathbf{y}_{i+1}),$$

where $\mathbf{y}_{i+1} = y_{i+1} \ldots y_{n_S}$ and $y_{n_S+1} = \emptyset$. The remaining terms will be proportional to

$$p(x, \mathbf{z}, \mathbf{y}_{i+1})$$

which will be $p(x, \mathbf{z})$ after the last family has been peeled.
To peel a family, say y_1, begin by specifying (A.5) as the following quadratic form in y_1:

$$p(y_1|x, \mathbf{z}, y_2 \ldots y_{n_S}) \propto \underbrace{\left(\frac{1}{\sigma_y^2} + \frac{n_p}{\sigma_z^2}\right)}_{a} y_1^2 - \underbrace{\left(\frac{2x}{\sigma_y^2} + \frac{2n_p \bar{\mathbf{z}}_1}{\sigma_z^2}\right)}_{b} y_1 + c \tag{A.6}$$

where

$$c = \left\{ \frac{x^2}{\sigma_x^2} + \frac{\sum_{i=2}^{n_S} (y_i - x)^2}{\sigma_y^2} + \sum_{i=2}^{n_S} \left(\frac{\sum_{j=1}^{n_p} (z_{ij} - y_i)^2}{\sigma_z^2} \right) \right\} + \frac{x^2}{\sigma_y^2} + \frac{\sum_j z_{1j}^2}{\sigma_z^2} \tag{A.7}$$

(A.6) can be rearranged in the form

$$\exp\left\{ \underbrace{a\left(y_1 - \frac{b}{2a}\right)^2}_{\propto p(y_1|x,\mathbf{z},\mathbf{y}_2)} + \underbrace{c - \frac{b^2}{4a}}_{\propto p(x,\mathbf{z},\mathbf{y}_2)} \right\}$$

$$p(x, \mathbf{z}, \mathbf{y}_2) \propto \exp\left\{ c - \frac{(x\sigma_z^2 + n_p \bar{z}_i \sigma_y^2)^2}{(\sigma_y^2 \sigma_z^2)(\sigma_z^2 + n_p \sigma_y^2)} \right\} \tag{A.8}$$

It follows, therefore, that:

$$p(x, \mathbf{z}) \propto \exp\left\{ \frac{x^2}{\sigma_x^2} + \frac{n_S x^2}{\sigma_y^2} + \sum_{i,j} \frac{z_{ij}^2}{\sigma_z^2} - \sum_{i=1}^{n_S} \frac{(x\sigma_z^2 + n_p \bar{z}_i \sigma_y^2)^2}{(\sigma_y^2 \sigma_z^2)(\sigma_z^2 + n_p \sigma_y^2)} \right\} \tag{A.9}$$

Rearranging (A.9) as a quadratic form in x:

$$p(x|\mathbf{z}) \propto \left(\frac{1}{\sigma_x^2} + \frac{n_S}{\sigma_y^2} - \frac{n_S \sigma_z^2}{\sigma_y^2(\sigma_z^2 + n_p \sigma_y^2)} \right) x^2 + \cdots \tag{A.10}$$

and therefore

$$\begin{aligned} \sigma_{x|\mathbf{z}}^2 &= \left(\frac{1}{\sigma_x^2} + \frac{n_S}{\sigma_y^2} - \frac{n_S \sigma_z^2}{\sigma_y^2(\sigma_z^2 + n_p \sigma_y^2)} \right)^{-1} \\ &= \left(\frac{1}{\sigma_x^2} + \frac{n_S n_p}{\sigma_z^2 + n_p \sigma_y^2} \right)^{-1} \\ &= \left(\frac{1}{\sigma_x^2} + \frac{N}{\sigma_z^2 + n_p \sigma_y^2} \right)^{-1} \end{aligned} \tag{A.11}$$

Equation (A.11) demonstrates that $\sigma_{x|\mathbf{z}}^2$ can be expressed a monotonically increasing function of n_p, and therefore the best case scenario is always to choose $n_s = N$ and $n_p = 1$. Qualitatively, the solution dictates that independent haplotypes are also preferable to additional family data.

Appendix B

Multimodality

One problem with data augmentation algorithms is that they can converge at a local, rather than global, optima. When the EM algorithm is employed, it is standard practice to run the algorithm many times from different initial parameter values. Estimates corresponding to the highest likelihood and are then selected. There is no guarantee that this strategy will work; the effectiveness is contingent upon the likelihood surface, which is determined by both the likelihood function and the data.

Studies using the EM algorithm for haplotype reconstruction typically restart the algorithm at least 100 times from different randomly generated haplotype frequency distributions. However, there have no studies to investigate whether this criterion is appropriate. One simple, yet useful, statistic that could be reported is the number of different modes were encountered when analyzing each data set.

Visualizing the likelihood surface for a given data set would clearly solve the problem. This appears impossible even for the simplest case involving two biallelic loci. (Recall that for L biallelic loci, there are 2^L haplotypes and therefore $2^L - 1$ free parameters.). However, it is possible to characterize the likelihood as a function of a single parameter when dealing with two biallelic loci. This can be done by recognizing that the only ambiguous genotype in this scenario will be double heterozygote, which can be resolved by two different phase configurations. The likelihood can then be be characterized in terms of a single parameter, α, which is defined as the proportion of double heterozygotes that are allocated to one of

	Locus 1	Locus 2	Log-Likelihood
y_1	0/0	0/0	$n_1 \log\left(\theta_{00}^2\right)$
y_2	0/0	0/1	$n_2 \log\left(2\theta_{00}\theta_{01}\right)$
y_3	0/0	1/1	$n_3 \log\left(\theta_{01}^2\right)$
y_4	0/1	0/0	$n_4 \log\left(2\theta_{00}\theta_{10}\right)$
y_5	0/1	0/1	$n_5 \log\left(2\theta_{00}\theta_{11} + 2\theta_{01}\theta_{10}\right)$
y_6	0/1	1/1	$n_6 \log\left(2\theta_{01}\theta_{11}\right)$
y_7	1/1	0/0	$n_7 \log\left(\theta_{10}^2\right)$
y_8	1/1	0/1	$n_8 \log\left(2\theta_{10}\theta_{11}\right)$
y_9	1/1	1/1	$n_9 \log\left(\theta_{11}^2\right)$

Table B.1: Components of the log-likelihood for a sample of phase-unknown genotypes. The data is expressed in terms of the counts $(n_1 \ldots n_9)$ for each of the 9 possible phase-unknown genotypes. The total log-likelihood is the sum over the final column.

the two possible phase configurations. For a given data set, it is straightforward to plot the log-likelihood against α.

B.1 Deriving the Likelihood

Consider two biallelic loci, where the alleles at each locus are denoted by 0 or 1. Each locus can therefore feature one of three possible genotypes, which are denoted $0/0$, $0/1$, $1/1$. Similarly, there are four haplotypes, $\mathbf{h} = (00, 01, 10, 00)$, with corresponding frequencies $\Theta = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$.

The likelihood for a given sample of size N is given by equation (2.3). The summation in (2.3) is over the number of individuals in the sample. It is straightforward to show that the counts for each of the 9 distinct phase-unknown genotypes are sufficient statistics. Table B.1 specifies each component of the likelihood function with data featuring genotype counts. The table illustrates that the only phase-unknown genotype where there is not a one-to-one correspondence between genotype and phase is y_5 (the double heterozy-

gote). Recall that when phase is observed, the maximum likelihood estimates are obtained through "gene counting", i.e. θ_i is proportional to the number of times h_i appears in the sample. The problem here is how to optimally allocate the n_5 double heterozygotes between (h_{00}, h_{11}) haplotypes that comprise the first phase possibility and the (h_{10}, h_{01}) haplotypes that comprise the second phase possibility.

Let α denote the proportion of double heterozygotes that are allocated to (h_{00}, h_{11}). Maximum likelihood estimates for Θ are given by fixed α:

$$\hat{\theta}_{00} = \frac{1}{2N}\{2n_1 + n_2 + n_4 + \alpha n_5\} \qquad \hat{\theta}_{11} = \tfrac{1}{2N}\{2n_9 + n_8 + n_6 + \alpha n_5\}$$

$$\hat{\theta}_{01} = \frac{1}{2N}\{2n_3 + n_2 + n_6 + (1-\alpha)n_5\} \quad \hat{\theta}_{10} = \tfrac{1}{2N}\{2n_9 + n_4 + n_8 + (1-\alpha)n_5\}$$

The likelihood can then be optimized by maximizing it with respect to α.

B.2 Results

For both empirical and simulated data sets based on HWE, such as the one in Figure B.1, multimodality was not indicated. It was, however, possible to induce multiple modes when using simulated data that deviated from HWE. In the scenario depicted in Figure B.2, all genotype counts other than double heterozygotes are held constant at 10. The number of double heterozygotes (n_5) is allowed to vary between 10 and 250 . It is interesting to note how the likelihood surface changes from unimodal, where all haplotypes are equally likely (i.e. $\alpha = .5$) to bimodal, where both modes are equally likely and each mode strongly favors one haplotype group to the other.

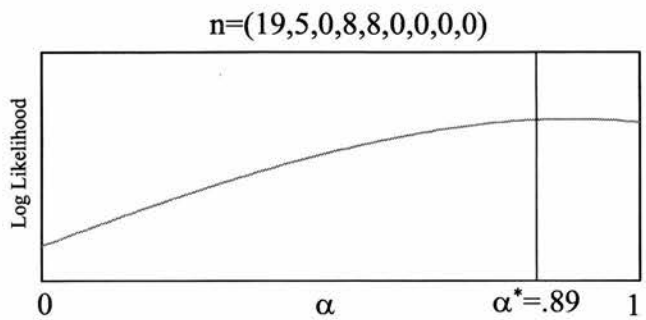


Figure B.1: Plot of log-likelihood against α for two-locus genotype counts (n) for *ldh1* and *Mdh* loci in mosquito data presented in Weir (1990).

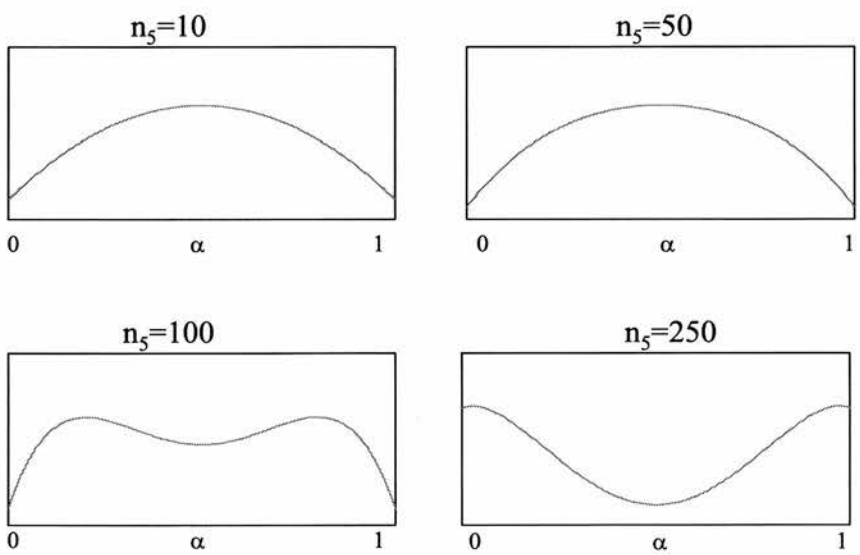


Figure B.2: Plot of log-likelihood against α for different counts of double heterozygotes (n_5). All other genotype counts are held constant at 10.

Appendix C

Estimating Haplotype Frequencies Using the EM Algorithm

This appendix provides three simple, yet instructive, examples to illustrate properties of the EM algorithm discussed in Section 2.2.1.

Throughout the appendix, we use notation introduced in Chapter 2. Additionally, in each of the three examples, we consider a two biallelic loci, where each allele on a given locus is denoted 1 and 2. The three possible genotypes that could be observed at given locus are denoted $1/1$, $1/2$, and $2/2$. The four haplotypes that can segregate at two loci are listed in Table C.1.

h_1	11
h_2	12
h_3	21
h_4	22

Table C.1: Haplotypes for two biallelic loci.

Phase configurations will be denoted as $h_i|h_j$. For example, $11|11$ denotes the homozygote for the 11 haplotype.

Example 1: Phase Is Observed

In the first example, we consider the sample of size three, given in Table C.2, where phase has been observed. Under the assumption of HWE, the likelihood for Θ follows a multino-

	Phase
z_1	11|11
z_2	12|21
z_3	22|22

Table C.2: A Sample of Three Phase-Known Genotypes Used In Example 1.

mial distribution, where each haplotype in each phase configuration can be regarded as a random sample from a probability vector, Θ. Hence, the complete data log-likelihood can be expressed as:

$$\log[p(\mathbf{z}|\Theta)] = \sum_{i=1}^{N}\sum_{j=1}^{M} n_{ij}\log\theta_j + \text{Constant}. \tag{C.1}$$

Maximum likelihood estimates for Θ are obtained through "gene counting". Specifically, calculating n_{ij}, which was defined in Chapter 2 as the number of times haplotype j appears in the phase configuration of individual i, can be trivially "counted":

$$\begin{array}{lll} n_{11}=2 & n_{21}=0 & n_{31}=0 \\ n_{12}=0 & n_{22}=1 & n_{32}=0 \\ n_{13}=0 & n_{23}=1 & n_{33}=0 \\ n_{14}=0 & n_{24}=0 & n_{34}=2 \end{array}$$

As shown in Table C.3, the maximum likelihood estimates of θ_i is the proportion of times h_i appears in the sample.

Example 2: Genotypes are Completely Informative for Phase

In this example, we consider the sample of three phase unknown genotypes given in Table C.4. When using the EM algorithm to calculate the maximum likelihood estimates for

θ_1	$\frac{2}{6}$
θ_2	$\frac{1}{6}$
θ_3	$\frac{1}{6}$
θ_4	$\frac{2}{6}$

Table C.3: Maximum Likelihood Estimates of Θ for Data Set Given in Example 1.

Θ, the latent (unobserved) data are the phase configurations. Note that when considering two biallelic loci, the only genotype that is not completely informative for phase is the one in which both loci are heterozygous. This phase configuration does not appear in the sample, and hence phase can be determined with certainty for each member in the sample. In principle, the appropriate phase configuration for each member of the sample can be deduced and maximum likelihood estimates for Θ can be obtained by calculating the complete data log-likelihood, as was done in the previous example. It is nonetheless instructive to consider the phase configurations latent and calculate one iteration of the EM algorithm. At the start of the EM algorithm, haplotype frequencies are assigned a random

	genotype	
y_1	$1/1$	$1/1$
y_2	$1/2$	$2/2$
y_3	$2/2$	$2/2$

Table C.4: A Sample of Three Genotypes which are Completely Informative For Phase.

value and normalized or, as shown in Table C.5, are assigned equal frequencies. From equation 2.1, $p(\mathbf{z}|\mathbf{y},\Theta^0)$ can be calculated:

$$p(z_1|y_1,\Theta^0) = \begin{cases} 1, & z_1 = 11|11 \\ 0, & \text{otherwise.} \end{cases} \tag{C.2}$$

$$p(z_1|y_1,\Theta^0) = \begin{cases} 1, & z_1 = 12|22 \\ 0, & \text{otherwise.} \end{cases} \tag{C.3}$$

	Θ^0
h_1	.25
h_2	.25
h_3	.25
h_4	.25

Table C.5: Initial Haplotype Frequency Estimates (Θ^0)

$$p(z_3|y_3,\Theta^0) = \begin{cases} 1, & z_1 = 22|22 \\ 0, & \text{otherwise.} \end{cases} \tag{C.4}$$

To compute the expected complete data log-likelihood, it is necessary to calculate the *expected* number of times each haplotype appears in a given individual:

$$\begin{array}{lll} E_{p(z_1|y_1,\Theta^0)}n_{11} = 2 & E_{p(z_2|y_2,\Theta^0)}n_{21} = 0 & E_{p(z_3|y_3,\Theta^0)}n_{31} = 0 \\ E_{p(z_1|y_1,\Theta^0)}n_{12} = 0 & E_{p(z_2|y_2,\Theta^0)}n_{22} = 1 & E_{p(z_3|y_3,\Theta^0)}n_{32} = 0 \\ E_{p(z_1|y_1,\Theta^0)}n_{13} = 0 & E_{p(z_2|y_2,\Theta^0)}n_{23} = 0 & E_{p(z_3|y_3,\Theta^0)}n_{33} = 0 \\ E_{p(z_1|y_1,\Theta^0)}n_{14} = 0 & E_{p(z_2|y_2,\Theta^0)}n_{24} = 1 & E_{p(z_3|y_3,\Theta^0)}n_{34} = 2 \end{array}$$

θ_i^0 is updated to θ_i^1 by calculating the *expected* proportion of times h_i appears in the sample. This is shown in Table C.6. Because phase configurations were determined directly from

θ_1	$\frac{2}{6}$
θ_2	$\frac{1}{6}$
θ_3	$\frac{0}{6}$
θ_4	$\frac{3}{6}$

Table C.6: Updated Haplotype Frequency Vector after One Iteration of the EM Algorithm (Θ^1).

genotype data, Θ^1 is independent of Θ^0, and that the next iteration will result in the same haplotype frequency vector, i.e. $\Theta^1 = \Theta^2$. Hence when genotypes are completely informative for phase, the EM algorithm will converge after one iteration, and the haplotype frequency estimates will be equivalent to those obtained by "gene counting".

Example 3: Sample Consisting Of Uninformative Genotypes

In this example, we consider the sample of three phase unknown genotypes given in Table C.7. Unlike the previous example, this sample contains a double heterozygote, and it is

	genotype	
y_1	1/1	1/1
y_2	1/2	1/2
y_3	1/1	1/2

Table C.7: The Sample of Three Genotypes Used in Example 3.

impossible to calculate maximum likelihood estimates by "gene counting". The EM algorithm is therefore employed.

As with the previous example, initial haplotype frequencies are assumed equally likely (Table C.5). The posterior distribution of the latent (phase) is then calculated:

$$p(z_1|y_1,\Theta^0) = \begin{cases} 1, & z_1 = 11|11 \\ 0, & \text{otherwise.} \end{cases} \tag{C.5}$$

$$p(z_2|y_2,\Theta^0) = \begin{cases} .5, & z_2 = 12|21 \\ .5, & z_2 = 11|22 \\ 0, & \text{otherwise.} \end{cases} \tag{C.6}$$

$$p(z_3|y_3,\Theta^0) = \begin{cases} 1, & z_1 = 11|12 \\ 0, & \text{otherwise.} \end{cases} \tag{C.7}$$

The expected complete data log-likelihood is derived by calculating the *expected* number of times each haplotype appears in a given individual:

$$\begin{array}{lll} E_{p(z_1|y_1,\Theta^0)}n_{11} = 2 & E_{p(z_2|y_2,\Theta^0)}n_{21} = .5 & E_{p(z_3|y_3,\Theta^0)}n_{31} = 1 \\ E_{p(z_1|y_1,\Theta^0)}n_{12} = 0 & E_{p(z_2|y_2,\Theta^0)}n_{22} = .5 & E_{p(z_3|y_3,\Theta^0)}n_{32} = 1 \\ E_{p(z_1|y_1,\Theta^0)}n_{13} = 0 & E_{p(z_2|y_2,\Theta^0)}n_{23} = .5 & E_{p(z_3|y_3,\Theta^0)}n_{33} = 0 \\ E_{p(z_1|y_1,\Theta^0)}n_{14} = 0 & E_{p(z_2|y_2,\Theta^0)}n_{24} = .5 & E_{p(z_3|y_3,\Theta^0)}n_{34} = 0 \end{array}$$

θ_i^0 is updated to θ_i^1 by calculating the *expected* proportion of times h_i appears in the sample. This is shown in Table C.8.

For the next iteration of the EM algorithm, the posterior distribution of the latent (phase)

θ_1	$\frac{3.5}{6} = .58$
θ_2	$\frac{1.5}{6} = .25$
θ_3	$\frac{.5}{6} = .083$
θ_4	$\frac{.5}{6} = .083$

Table C.8: Updated Haplotype Frequency Vector after One Iteration of the EM Algorithm (Θ^1).

is again calculated:

$$p(z_1|y_1,\Theta^1) = \begin{cases} 1, & z_1 = 11|11 \\ 0, & \text{otherwise.} \end{cases} \tag{C.8}$$

$$p(z_2|y_2,\Theta^1) = \begin{cases} .3, & z_2 = 12|21 \\ .7, & z_2 = 11|22 \\ 0, & \text{otherwise.} \end{cases} \tag{C.9}$$

$$p(z_3|y_3,\Theta^1) = \begin{cases} 1, & z_1 = 11|12 \\ 0, & \text{otherwise.} \end{cases} \tag{C.10}$$

As before, the expected complete data log-likelihood id derived by calculating the *expected* number of times each haplotype appears in a given individual:

$$\begin{array}{lll} E_{p(z_1|y_1,\Theta^1)}n_{11} = 2 & E_{p(z_2|y_2,\Theta^1)}n_{21} = .7 & E_{p(z_3|y_3,\Theta^1)}n_{31} = 1 \\ E_{p(z_1|y_1,\Theta^1)}n_{12} = 0 & E_{p(z_2|y_2,\Theta^1)}n_{22} = .3 & E_{p(z_3|y_3,\Theta^1)}n_{32} = 1 \\ E_{p(z_1|y_1,\Theta^1)}n_{13} = 0 & E_{p(z_2|y_2,\Theta^1)}n_{23} = .3 & E_{p(z_3|y_3,\Theta^1)}n_{33} = 0 \\ E_{p(z_1|y_1,\Theta^1)}n_{14} = 0 & E_{p(z_2|y_2,\Theta^1)}n_{24} = .7 & E_{p(z_3|y_3,\Theta^1)}n_{34} = 0 \end{array}$$

θ_i^1 is updated to θ_i^2 by calculating the *expected* proportion of times h_i appears in the sample. This is shown in Table C.9. The process is repeated until $\Theta^k \approx \Theta^{k+1}$. We can clearly see after the first iteration that h_1 will have the highest frequency estimate. This is to be expected because it is the most prevalent in the completely informative genotypes.

θ_1	$\frac{3.7}{6} = .62$
θ_2	$\frac{1.3}{6} = .22$
θ_3	$\frac{.3}{6} = .05$
θ_4	$\frac{.7}{6} = .11$

Table C.9: Updated Haplotype Frequency Vector after the Second Iteration of the EM Algorithm (Θ^2).

Glossary

Allele One of the alternative versions of a polymorphism.

Association Analysis Any statistical method that tests whether a certain allele or haplotype is found significantly more frequently in a group of affected individuals than in a group of unrelated controls.

Genotype The pair of alleles at a given locus in an individual.

Haplotype Any sequence of alleles that are linked on a chromosome.

Identical By Descent (IBD) When two alleles are inherited from a shared ancestor.

Linkage The association of alleles on the same chromosome. During meiosis, alleles that are linked are transmitted more frequently than chance would allow.

Linkage Analysis A statistical method that tests for the coinheritence of genetic markers with biological traits or other markers within families.

Linkage Disequilibrium (LD) The correlation between polymorphisms that arises because variants share a joint population ancestry.

Linkage Equilibrium (LE) Refers to the phenomenon when allele frequencies are independent.

Phase Pair of haplotypes in an individual.

Bibliography

Abecasis, G. R., Cherny, S., Cookson, W., and Cardon, L. (2002). Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, 30:97–101.

Abecasis, G. R. and Wigginton, J. E. (2005). Handling Marker-Marker Linkage Disequilibrium: Pedigree Analysis with Clustered Markers. *Am. J. Hum. Genet.*, 77:754–767.

Becker, T. and Knapp, M. (2002). Efficiency of Haplotype Frequency Estimation when Nuclear Family Information is Included. *Hum. Hered.*, 54:45–53.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Gen.*, 33:228–237.

Cannings, C., Thompson, E., and Skolnick, M. (1978). Probability Functions on Complex Pedigrees. *Adv. Appl. Prob.*, 10:26–61.

Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2003). Detecting Disease Associations due to Linkage Disequilibrium using Haplotype Tags: A Class of Tests and the Determinants of Statistical Power. *Human Heredity*, 56:18–31.

Clayton, D. (2002). Snphap: A program for estimating frequencies of large haplotypes of snps. *www.gene.cimr.cam.ac.uk*.

Elston, R. and Stewart, J. (1971). A General Model for the Genetic Analysis of Pedigree Data. *Hum. Hered.*, 21:523–542.

Excoffier, L. and Slatkin, M. (1995). Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. *Mol. Biol. Evol.*, 12:921–927.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D., and Schork, N. J. (2001). Genetic Analysis of Case/Control Data Using Haplotype Frequencies: Application to APOE Locus Variation and Alzheimer's Disease. *Gen. Res.*, 11:143–151.

Fallin, D. and Shorck, N. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am. J. Hum. Genet.*, 67:947–959.

Fan, R. and Knapp, M. (2003). Genome association studies of complex disease by case control designs. *Am. J. Hum. Genet.*, 53:850–868.

Fishelson, M., Dovgolevsky, N., and Geiger, D. (2005). Maximum Likelihood Haplotyping for General Pedigrees. *Human Heredity*, 59:41–60.

Fishelson, M. and Geiger, D. (2002). Exact Genetic Linkage Computations for General Pedigrees. *Bioinformatics*, 18:189–198.

Grapes, L., Dekkers, J., Rothschild, M., and Fernando, R. (2004). Comparing Linkage Disequilibrium-Based Methods for Fine Mapping of Quantitative Trait Loci. *Genetics*, 166:1561–1570.

Hawley, M. and Kidd, K. (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.*, 86:409–411.

Heifetz, E., Fulton, J., Sullivan, N., Zhao, H., Dekkers, J., and Soller, M. (2005). Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics*, 171:1173–1181.

Hill, W. G. (1974). Estimation of Linkage Disequilibrium in Randomly Mating Populations. *Hered*, 33:229–239.

Hull, J., Ackerman, H., Isles, K., Usen, S., Pinder, M., Thomson, A., and Kwiatkowski, D. (2001). Unusual Haplotypic Structure of IL8, a Susceptibility Locus for a Common Respiratory Virus. *Am. J. Hum. Genet.*, 69:413–419.

Ito, T., Chiku, S., Inohue, E., Tomita, M., Morisaki, T., Morisaki, H., and Kamatabi, N. (2003). Estimation of Haplotype Frequencies, LD Measures and Combination of Haplotype Copies in Each Pool by Use of Pooled DNA Data. *PNAS*, 72:384–398.

Jordan, M. I. (2003). *An Introduction to Probabilistic Graphical Models*. Draft.

Kirk, K. M. and Cardon, L. R. (2002). The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur. J. Hum. Genet.*, 10:616–622.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308:385–389.

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

Lange, K. and Boehnke, M. (1983). Estensions to Pedigree Analysis v Opimal Calculations of Mendelian Likelihoods. *Hum Hered*, 33:291–301.

Lange, K. and Elston, R. (1975). Extensions to Pedigree Analysis. Likelihood Calculations for Simple and Complex Pedigrees. *Hum Hered*, 25:95–105.

Lauritzen, S. L. and Sheehan, N. A. (2003). Graphical Models for Genetic Analyses. *Statistical Science*, 18:489–514.

Lee, S. H. and van der Werf, J. H. (2004). The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. *Genet. Sel. Evol.*, 36:145–160.

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., and Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to suseptibility to common disease. *Nat. Gen.*, 33:177–180.

Long, A. D. and Langely, C. H. (1999). The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. *Genet. Res*, 13:720–730.

Lu, X., Niu, T., and Liu, J. S. (2003). Haplotype Information and Linkage Disequilibrium Mapping for Single Nucleotide Polymorphisms. *Genet. Res*, 13:2112–2117.

Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsaker, I., and Goddard, M. E. (2002). Fine Mapping of a Quantitative Trait Locus for Twinning Rate Using Combined Linkage and Linkage Disequilibirum Mapping. *Genetics*, 161:373–379.

Morris, A., Whittaker, J., and Balding, D. (2002). Fine-Scale Mapping of Disease Loci via Shattered Coalescent Modeling of Genealogies. *Am. J. Hum. Genet.*, 70:686–707.

Nejati-Javaremi, A. and Smith, C. (1996). Assigning Linkage Haplotypes From Parent and Progeny Genotypes. *Genetics*, 142:1363–1369.

Nielsen, D. M., Ehm, M. G., Zaykin, D. V., and Weir, B. S. (2004). Effect of tow and three locus linkage disequilibrium on the power to detect marker phenotype associations. *Genetics*, 168:1029 – 1040.

Niu, T. (2004). Algorithms for Inferring Haplotypes. *Genet. Epidemiol.*, 27:224–247.

Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002). Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. *Am. J. Hum. Genet.*, 70:157–169.

O'Connell, J. R. (2000). Zero-Recombinant Haplotyping: Applications to Fine Mapping Using SNPs. *Genet. Epidemiol.*, 19:S582–S587.

Perez-Enciso, M. (2003). Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics*, 163:1497–1510.

Qian, D. and Beckmann, L. (2002). Minimum-Recombinant Haplotyping in Pedigrees. *Am. J. Hum. Genet.*, 70:1434–1445.

Qin, Z. S., Niu, T., and Liu, J. S. (2002). Partial-ligation-expectation-maximization for haplotype inference with single nucleotide polymorphisms. *Am. J. Hum. Genet.*, 71:1242–1247.

Rohde, K. and Fuerst, R. (2001). Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Hum Mutat*, 17:289–295.

Schaid, D. J. (2002). Relative Efficiency of Ambiguous vs. Directly Measured Haplotype Frequencies. *Genet. Epidemiol.*, 23:426–443.

Schaid, D. J., McDonnell, S. K., Wang, L., Cunningham, J. M., and Thibodeau, S. N. (2002). Caution on Pedigree Haplotype Inference with Software that Assumes Linkage Equilibrium. *Am. J. Hum. Genet.*, 71:992–995.

Schouten, M. T., Williams, C. K. I., and Haley, C. S. (2005). The impact of using related individuals for haplotype reconstruction in population studies. *Genetics*, 171:1–10.

Sobel, E. and Lange, K. (1996). Descent Graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.*, 58:1323–2337.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc. B.*, 62:605–655.

Stephens, M. and Donnelly, P. (2003). A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *Am. J. Hum. Genet.*, 73:1162–1169.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *Am. J. Hum. Genet.*, 68:178–989.

Terwilliger, J. and Ott, J. (1994). *Handbook of Human Linkage Analysis*. John Hopkins University Press.

Thomas, A. (2003). Accelerated gene counting for haplotype frequency estimation. *Annals of Hum Genet*, 67:608–612.

Thomas, S. C. and Hill, W. G. (2002). Sibship reconstruction in hierarchical population structures using Markov Chain Monte Carlo techniques. *Genet. Res. Camb.*, 79:227–234.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A., Kidd, J., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A., P.Moral, Krings, M., Paabo, S., Watson, E., Risch, N., Jenkins, T., and Kidd, K. (1996). Global Patterns of Linkage Disequilibrium at the CD4 Locus and Modern Human Origins. *Sci*, 271:1380–1387.

Wang, W. Y., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-Wide association Studies: Theoretical and Practical Concerns. *Nat Rev Genet*, 6:109–117.

Weir, B. S. (1990). *Genetic Data Analysis*. Sinauer Associates.

Weiss, K. M. and Clark, A. G. (2002). Linkage disequilibirum and the mapping of complex traits. *Trends. Genet.*, 18:19–24.

Yang, Y., Zhang, J., Hoh, J., Matsuda, F., Xu, P., Lathrop, M., and Ott, J. (2003). Efficiency of SNP haplotype estimation from pooled DNA. *PNAS*, 100:7225–7230.

Zaykin, D. V., Westfall, P. H., and Ehm, M. G. (2001). Testing associations of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.*, 53:79–91.

Zhang, K., Sun, F., and Zhao, H. (2005). Haplore: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics*, 21:90–103.

Zhang, S., Pakstis, A. J., Kidd, K. K., and Zhao, H. (2001). The complex interplay among factors that influence allelic association. *Am. J. Hum. Genet.*, 69:906–912.

Zhao, H., Pfeiffer, R., and Gail, M. H. (2003). Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, 4:171–178.

Zondervan, K. T. and Cardon, L. R. (2003). The complex interplay among factors that influence allelic association. *Nature Reviews*, 4:89–100.