Approaches To Estimation Of Haplotype Frequencies And Haplotype-Trait Associations

Item Type	dissertation
Authors	Li, Xiaohong
DOI	10.7275/5648858
Download date	2024-12-11 04:28:37
Link to Item	https://hdl.handle.net/20.500.14394/16862

APPROACHES TO ESTIMATION OF HAPLOTYPE FREQUENCIES AND HAPLOTYPE-TRAIT ASSOCIATIONS

A Dissertation Presented

by

XIAOHONG LI

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2009

Public Health

APPROACHES TO ESTIMATION OF HAPLOTYPE FREQUENCIES AND HAPLOTYPE-TRAIT ASSOCIATIONS

A Dissertation Presented

by

XIAOHONG LI

Approved as to style and content by:

_____
Andrea S. Foulkes, Chair

_____
John P. Buonaccorsi, Member

_____
John Staudenmayer, Member

_____
Rongheng Lin, Member

_____
Michael E. Begay, Department Chair
Public Health

To Jianbin, Chenchen, my mother and father

ACKNOWLEDGMENTS

ABSTRACT

APPROACHES TO ESTIMATION OF HAPLOTYPE FREQUENCIES AND HAPLOTYPE-TRAIT ASSOCIATIONS

FEBRUARY 2009

XIAOHONG LI

B.E., NORTHERN JIAOTONG UNIVERSITY

M.E., INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Andrea S. Foulkes

Characterizing the genetic contributors to complex disease traits will inevitably require consideration of haplotypic phase, the specific alignment of alleles on a single homologous chromosome. In population based studies, however, phase is generally unobservable as standard genotyping techniques provide investigators only with data on unphased genotypes. Several statistical methods have been described for estimating haplotype frequencies and their association with a trait in the context of phase ambiguity. These methods are limited, however, to diploid populations in which individuals have exactly two homologous chromosomes each and are thus not suitable for more general infectious disease settings. Specifically, in the context of Malaria and HIV, the number of infections is also unknown. In addition, for both diploid and non-diploid settings, the challenge of high-dimensionality and an unknown model of asso-

ciation remains. Our research includes: (1) extending the expectation-maximization approach of Excoffier and Slatkin to address the challenges of unobservable phase and the unknown numbers of infections; (2) extending the method of Lake et al. to estimate simultaneously both haplotype frequencies and the haplotype-trait associations in the non-diploid settings; and (3) application of two Bayesian approaches to the mixed modeling framework with unobservable cluster (haploype) identifiers, to address the challenges associated with high-dimensional data. Simulation studies are presented as well as applications to data arising from a cohort of children multiply infected with Malaria and a cohort of HIV infected individuals at risk for anti-retroviral associated dyslipidemia. This research is joint work with Drs. S.M. Rich, R.M. Yucel, J. Staudenmayer and A.S. Foulkes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Characterizing the genetic contributors to complex disease traits will inevitably require consideration of haplotypic phase, the specific alignment of alleles on a single homologous chromosome. In population based studies, however, phase is generally unobservable as standard genotyping techniques only provide data on unphased genotypes. Several statistical methods have been described for estimating haplotype frequencies [4, 8, 40, 28, 34] and their association with a trait [25, 27] in the context of phase ambiguity. These methods are limited, however, to diploid populations in which individuals have exactly two homologous chromosomes each and are thus not suitable for more general infectious disease settings. Specifically, in the context of Malaria and HIV, the number of infections is unknown. In addition, for both diploid and non-diploid settings, the challenge of high-dimensionality and an unknown model of association remains. Our research tries to address these problems by developing efficient methods that can be used in more general settings

First, we extend the Expectation-Maximization(EM) approach of Excoffier and Slatkin [8] to address the challenges of unobservable phase and the unknown numbers of infections. Excoffier and Slatkin [8] uses the EM algorithm to find the haplotype frequencies maximizing the likelihood of observed sample of genotypes. For the human genetic setting where the number of clones is always 2, this approach is straightforward and effective. However, this method can only be applied in the diploid population and thus not suitable for more general settings like Malaria. The approach we used to handle the difficulty caused by the unknown number of clones (i.e. the number of

infections) is to impose a probability distribution on this number. More specifically, three distributional assumptions are considered: fixed, unconditional Poisson (UP) and conditional Poisson (CP). Our method can be reduced to the approach of [8] when the number of infections is fixed at 2. The method we proposed addresses directly the challenges of unobservable phase and the unknown numbers of infections, while providing a computationally efficient framework that accommodates multiple genetic loci.

Second, we extend the method of Lake et al. [25] to simultaneously estimate both haplotype frequencies and the haplotype-trait associations in the non-diploid settings. Lake et al. [25] uses EM algorithm to address the unknown haplotypic phase and provides a comprehensive framework for simultaneous estimation of population haplotype frequencies and haplotype-trait associations under generalized linear model framework. This method, again, is limited to the diploid population and not suitable for general settings. Our method can deal with variable number of infections by modeling this number. Similarly to the approach used in the first work, three assumptions on the number of infections, including fixed, conditional Poisson(CP) and Semi-parametric, are considered.

Third, We apply two Bayesian approaches to the mixed modeling framework with unobservable cluster identifiers. Mixed modeling is a useful approach for characterizing haplotype-trait associations in the context of population-based association studies of unrelated individuals. In this case, clusters are often defined as groups of genetically similar individuals, for example, the individuals who carry a common pair of haplotypes. The problem of this approach is that haplotypic phase is unobservable. Therefore, the cluster identifier is ambiguous and general estimation methods for mixed model cannot be directly applied. In a recent manuscript, Foulkes and Yucel [11] describe an Expectation Conditional Maximization (ECM) approach to account for uncertainty in the cluster identifiers arising from unobserved haplotypic

phase. The downside of that method is that it is computational unfeasible when the number of ambiguous individuals within one cluster is large. In the method we proposed, two prior distributions are assumed for cluster effects. We first consider a single normal prior and then we relax the strict normality assumption and assume random cluster effects arise from a discrete mixture distribution with a Dirichlet process prior. Gibbs sampler are used for iteratively arriving at estimates. In order to account for the unknown cluster identifier, we propose to impute the cluster membership for each individual at the beginning of each iteration. This approach have a marked computational advantage over the ECM approach.

The next three chapters of this thesis describe these three parts of research in detail. Chapter 5 summarizes all these work and discusses future work.

CHAPTER 2

AN EXPECTATION MAXIMIZATION APPROACH TO ESTIMATE MALARIA HAPLOTYPE FREQUENCIES IN MULTIPLY INFECTED CHILDREN

SUMMARY: Characterizing genetic variability in the human pathogenic *Plasmodium* species, the group of parasites that cause Malaria, may have broad global health implications. Specifically, discerning the combinations of mutations that lead to viable parasites and the population level frequencies of these clonal sequences will allow for targeted vaccine development and individualized treatment choices. This presents an analytical challenge, however, since haplotypic phase (i.e. the alignment of bases on a single DNA strand) is generally unobservable in multiply infected individuals. This manuscript describes an expectation maximization (EM) approach to maximum likelihood estimation of haplotype frequencies in this missing data setting. The approach is applied to a cohort of N=341 malaria infected children in Uganda, Cameroon and Sudan to characterize regional differences. A simulation study is also presented to characterize method performance and assess sensitivity to distributional assumptions.

2.1 Introduction

Malaria continues to be a grave public health concern with an estimated 1-3 million associated deaths per year [3], the estimated majority ($> 75\%$) of which occur in African children under the age of 5 [17]. Malaria is an infectious disease caused by a group of parasites called the human pathogenic *Plasmodium* species. Development of effective vaccines and appropriate treatment intervention strategies will inevitably require characterizing the genetic variability of the parasites both within and across

geographic regions. This, however, presents an analytical challenge due to the large number of polymorphic sites in the parasite genome and the presence of multiple clonal species within an individual host.

Vaccine development efforts generally target the cellular adhesion molecule, circumsporozoite protein (CSP). This protein is expressed by the parasite and facilitates adhesion to hepatocytes (liver cells) in the human host [44, 21]. It is in the host liver that the parasite multiplies and eventually differentiates into merozoites (daughter cells) that infect red blood cells, leading to the common symptoms of malaria. In this investigation, we focus on one polymorphic region (csp-th3) within CSP that encodes a T-cell epitope. The data motivating our research arise from a cross-sectional study of $n = 341$ malaria infected children from three African nations: Uganda, Cameroon and Sudan. The genetic region consists of 12 loci, of which 10 are variable in our cohort.

In endemic areas such as those we consider, each infected child commonly carries between 1 and 5 different clonal species. Here the term clone is used to indicate a genetically unique parasite. Malarial parasites are typically transmitted from one human host to another via a mosquito vector. Sexual differentiation into male and female gametes occurs in the mosquito and sporozoites with a *single* chromosomal copy of a given malaria parasite are injected into the human host at the time a blood meal is taken. Multiple infections generally arise due to (1) multiple parasites being transmitted in a single blood meal and/or (2) multiple mosquitos taking blood meals from a single individual.

The parasite genome is comprised of a sequence of bases represented by the four letters A, C, T and G. Locations within this sequence that are variable across the general population are termed single nucleotide polymorphism (SNP) loci. Data across multiple SNP loci are referred to as multi-locus genotypes. The process of individually genotyping each of the clonal sequences infecting a single human host

is both laborious and expensive. For this reason, population assays are typically generated that report the set of bases present at each locus on the parasite genome within an individual. That is, for example, suppose an individual is infected with two clones: the first clone has an A at locus one and the second clone has a G at this locus. In this case, the observed genotype for this individual is given by A/G. Notably, if the individual were additionally infected with a third clone that had a G at this locus, then the observed genotype would be identical (A/G). That is, the number of copies of a base within an individual at a given location is not observed.

As a result of the assay utilized, the alignment of bases on a given clone (commonly referred to as haplotypic phase) is not observable. For example, suppose the parasite population within an individual is sequenced and the two bases A and T are observed at locus 1 and the bases G and C are observed at locus 2. That is, the genotype across these two loci is given by $(A/T, G/C)$. In this simple case, there are four possible alignments of bases on a single clone: $h_1 = (A, G), h_2 = (A, C), h_3 = (T, G)$ and $h_4 = (T, C)$. The precise combination of these haplotypes within a person is not observable. For example, the true haplotypes could be (h_1, h_4) or (h_2, h_3). Moreover, since the number of clonal sequences within each person is also unobserved, the number of copies of each haplotype is unknown. For example, the true haplotype combination could also be (h_1, h_4, h_4) or (h_1, h_1, h_4, h_4) and depends on whether the individual has 2, 3 or 4 infections. This is described further in Section 2.2 after additional notation is outlined. Knowledge of haplotypic phase is particularly relevant to the development of appropriate treatment and vaccine strategies that target large (i.e. multi-site) regions of the parasite's genome.

Several statistical approaches to inferring haplotypes from unphased genotype data have been proposed in the context of human genetic investigations. These include Clark's algorithm [4], an EM-type algorithm [8] and more recently several Bayesian approaches, including those of [40], [28] and [34]. Notably, these approaches

6

were developed for diploid populations in which each individual has exactly two chromosomes. In our setting, the number of clones (infections) is variable. That is, each individual has multiple infections and the number of infections is not observed, rendering the aforementioned approaches unsuitable. [20] derived a hill-climbing approach for this multiple clone setting; however, it is computationally intensive when the number of variable loci is greater than two. This manuscript presents a novel extension of the EM algorithm of [8] for haplotype reconstruction in this non-diploid setting.

Our method addresses directly the challenges of unobservable phase and the unknown numbers of infections, while providing a computationally efficient framework that accommodates multiple genetic loci. This approach uses the general framework of the EM algorithm, formalized by [5]. Since in our setting the number of clones (i.e. the number of infections) is not known, we impose a probability distribution on this number. Three (3) distributional assumptions are considered: fixed, unconditional Poisson (UP) and conditional Poisson (CP). Our method reduces to the approach of [8] when the number of infections is fixed at 2, which is straightforward to implement using the `haplo.em()` function in the R package `haplo.stats`. By framing unobservable haplotypic phase as a missing data problem, we derive an EM approach that is intuitively appealing and guarantees an increase in the likelihood at each iteration.

Notably, our approach is also closely related to methods described for the estimation of population haplotype frequencies with pooled DNA samples [42, 22, 35]. Specifically, we propose a generalization of the EM approaches of [42] and [22] in which the number of clones is fixed at $2 * M$ where M is the number of samples in the pool. In Section 2.2.2 of this manuscript, we describe a similar EM approach assuming a fixed number (c) of infections and extend this in Section 2.2.3 to the more general setting of a variable number of infections.

We begin in the following section by describing our modeling assumptions and the EM procedure. In Section 2.3 we present a simulation study to characterize method performance. Finally, application of the EM approach to a cross-sectional study of $n = 341$ children in equatorial Africa is presented in Section 2.4 to characterize regional differences in haplotype frequencies.

2.2 Methods

Estimating haplotype frequencies in a population of multiple infected individuals requires making a distributional assumption about the number of infections. We assume that each infection is independent. While multiple infections can arise with a single mosquito bite, we are unable to distinguish between this occurrence and the occurrence of multiple blood meals taken by different mosquitos. In Section 2.2.1 we begin by outlining our notation. The approach presented in Section 2.2.2 assumes the number of infections is fixed at a constant, $c > 0$. While this assumption is probably not reasonable here, it represents a natural extension of the method of [8] and is useful for comparison. It may also be useful for other settings, as described in Section 2.5. In Section 2.2.3 we extend this approach further by allowing for a variable number of clonal infections across our population. Specifically, both the unconditional Poisson (UP) and conditional Poisson (CP) assumptions are considered.

2.2.1 Notation

Let H_i represent the combination of haplotypes and G_i be the unphased multi-locus genotype for the ith individual, $i = 1, \ldots, n$. In our setting, G_i is known for each individual, while H_i is unknown. There are generally multiple values of H_i consistent with G_i. In a diploid population, H_i consists of a pair of haplotypes; in general, however, the number of elements of H_i will vary since a combination of 1 or more

haplotypes may be consistent with the observed genotype G_i. That is, the number of clones (C) in a combination may vary.

Let $\mathcal{S}(G_i)$ be the set of all haplotype combinations that are consistent with genotype G_i. Further let $h_1, \ldots, h_K$ denote the K possible haplotypes over all observed individuals, let θ_k be the population frequency of haplotype h_k and define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. In general, interest lies in estimating $\boldsymbol{\theta}$. We propose a novel application of the EM algorithm to arrive at a maximum likelihood estimate of $\boldsymbol{\theta}$.

Consider for example the simple case of two biallelic loci: the first locus is either A or a and the second locus is B or b. There are 4 possible haplotypes across all individuals: $h_1 = (A, B)$, $h_2 = (A, b)$, $h_3 = (a, B)$, and $h_4 = (a, b)$. For the observed genotype $G_i = (A/a, B/b)$, the specific haplotype combination H_i is unknown, although there are many combinations consistent with this genotype. The set of possible haplotype combinations is given by $\mathcal{S}(G_i) = \{(h_1, h_4), (h_1, h_4, h_4), (h_2, h_3), (h_2, h_3, h_4), \ldots\}$ and the true (unobserved) haplotype $H_i \in \mathcal{S}(G_i)$. Note that since heterozygosity is observed at 2 loci for this genotype, we know there are at least 2 clonal sequences. Further, multiple copies of the same haplotype may be present in a single individual. Thus, for example it is possible that $H_i = (h_1, h_1, h_4)$. A visual representation of this unobservable data is given in Figure 2.2.1.

Finally, we define δ_{ik} to be the number of copies of haplotype h_k present in the haplotype combination H_i, $k = 1, \ldots, K$. For example δ_{i1} is the number of copies of h_1 in the haplotype combination H_i and is given by:

$$
\delta_{i1} = \begin{cases} 1 & \text{if } H_i = (h_1, h_4) \\ 1 & \text{if } H_i = (h_1, h_4, h_4) \\ 2 & \text{if } H_i = (h_1, h_1, h_4) \\ \vdots & \end{cases}
$$

Figure 2.1. Unobservable Haplotype Combinations

All 29 possible haplotype combinations for clone number $c \leq 3$ are illustrated. The combinations consistent with the genotype $AaBb$ are represented by shaded boxes.

2.2.2 Estimation Assuming Known (Fixed) Number of Infections

Suppose there are exactly $C > 0$ clones present in each blood sample. That is, assume each individual has exactly C infections, where C is some positive integer. In this case we have $\sum_{k=1}^{K} \delta_{ik} = C$ for $i = 1, \ldots, n$ and we can write the haplotype set probabilities, $Pr(H_i|\boldsymbol{\theta})$ by:

$$Pr(H_i|\boldsymbol{\theta}) = \frac{C!}{\delta_{i1}! \ldots \delta_{iK}!} \prod_{k=1}^{K} \theta_k^{\delta_{ik}} \tag{2.1}$$

Application of the EM algorithm typically proceeds by defining the complete data, denoted $\mathbf{X}_{com}$ in terms of the observed and missing components. In our setting, the observed data $\mathbf{X}_{obs}$ is the genotype G_i for each individual and the complete data is both the genotype and specific haplotype combination H_i for each individual. The complete data likelihood is given by:

$$L(\boldsymbol{\theta}|\mathbf{X}_{com}) = \prod_{i=1}^{n} Pr(H_i|\boldsymbol{\theta}) \tag{2.2}$$

where $Pr(H_i|\boldsymbol{\theta})$ is as defined in Equation 2.1.

The E-step involves calculating the expectation of the complete data log likelihood conditional on the observed data $G_i, i = 1, \ldots, n$. This conditional expectation is given by:

$$E\left[\log L(\boldsymbol{\theta}|\mathbf{X}_{com})\right] = \sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\boldsymbol{\theta}) \log Pr(H_i|\boldsymbol{\theta}) \tag{2.3}$$

where $p_{iH_i}(\boldsymbol{\theta})$ is the posterior probability of H_i given G_i. A formulation of this posterior probability is given by:

$$p_{iH_i}(\boldsymbol{\theta}) = \frac{Pr(H_i|\boldsymbol{\theta})}{\sum_{H_i \in \mathcal{S}(G_i)} Pr(H_i|\boldsymbol{\theta})} \tag{2.4}$$

11

The M-step of the EM algorithm maximizes the conditional expectation of the complete data log likelihood given in Equation 2.3. Let $\boldsymbol{\theta}^{(t)}$ be the estimate of $\boldsymbol{\theta}$ derived from the tth iteration of the EM algorithm. The $(t+1)^{th}$ estimate of $\boldsymbol{\theta}$ can be obtained by finding the root for the following equation.

$$
\begin{aligned}
\frac{\partial E\left[\log L(\boldsymbol{\theta}|\mathbf{X}_{com})\right]}{\partial \theta_k} &= \sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\boldsymbol{\theta}}^{(t)}) \partial \log Pr(H_i|\boldsymbol{\theta})/\partial \theta_i \\
&= \sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\boldsymbol{\theta}}^{(t)})(\frac{\delta_{ik}}{\theta_k} - \frac{\delta_{iK}}{\theta_K}) = 0
\end{aligned}
\tag{2.5}
$$

Resulting closed form solutions for $\hat{\theta}_k$ are given by:

$$
\widehat{\theta}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\boldsymbol{\theta}}^{(t)}) \delta_{ik}}{nC}
\tag{2.6}
$$

2.2.3 Estimation Assuming Unknown (Variable) Numbers of Infections

In Section 2.2.2 we assume that the number of clones (infections) is fixed; however, in general each individual may have a different number of clones. In this section we relax this assumption and instead assume a probability distribution on the number of infections per individual.

2.2.3.1 Unconditional Poisson Assumption

We begin by assuming the number of infections has a Poisson distribution with probability density given by:

$$
\phi_c(\lambda) = e^{-\lambda}(\lambda^c/c!)
\tag{2.7}
$$

Note this density is a function of the rate parameter, λ which can be interpreted as the average number of infections across the population. Equation 2.1 for the haplotype combination probabilities is now replaced by:

$$Pr(H_i|\Phi) = \phi_{c_i}(\lambda)\frac{c_i!}{\delta_{i1}!\ldots\delta_{iK}!}\prod_{k=1}^{K}\theta_k^{\delta_{ik}} \tag{2.8}$$

where $\Phi = (\boldsymbol{\theta}, \lambda)$. Here c_i is the number of clones in the set H_i and can vary across i.

Let n_0 be the number of uninfected individuals. Without loss of generality, we assume individuals are ordered so that $i = 1, \ldots, (n - n_0)$ represent individuals with at least one detected infection and $i = (n - n_0 + 1), \ldots, n$ are uninfected individuals. The complete data likelihood for this setting is given by:

$$L(\Phi|\mathbf{X}_{com}) = Pr(c = 0)^{n_0}\prod_{i=1}^{n-n_0}Pr(H_i|\Phi) \tag{2.9}$$

where $P_r(c = 0) = e^{-\lambda}$. The expected conditional log likelihood is:

$$E\left[\log L(\Phi|\mathbf{X}_{com})\right] = -n_0\lambda + \sum_{i=1}^{n-n_0}\sum_{H_i\in\mathcal{S}(G_i)}p_{iH_i}(\Phi)\log Pr(H_i|\Phi) \tag{2.10}$$

where $p_{iH_i}(\Phi)$ is defined in Equation 2.4 with $\boldsymbol{\theta}$ replaced by Φ.

Note that in our cross-sectional investigation of infected children n_0 is set equal to 0. In general, random sampling of children would allow for incorporating the number of individuals with no infection. Setting n_0 equal to 0 in the cross-sectional setting leads to overestimation of λ. For this reason, we recommend use of the conditional Poisson approach described in Section 2.2.3.2 when sampling consists only of infected children. Further discussion of this point is provided in Section 2.4.

Estimation of $\boldsymbol{\theta}$ is similar to the setting in which C is fixed. Resulting closed form solutions for $\hat{\theta}_k$ are:

$$\widehat{\theta}_k^{(t+1)} = \frac{\sum_{i=1}^{n-n_0} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik}}{\sum_{i=1}^{n-n_0} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) c_i} \tag{2.11}$$

where $\widehat{\Phi}^{(t)}$ is the estimate of Φ derived from the kth iteration of the EM algorithm.

Estimation of λ is achieved by solving:

$$\begin{aligned}
\frac{\partial E\left[\log L(\Phi|\mathbf{X}_{com})\right]}{\partial \lambda} &= -n_0 + \sum_{i=1}^{n-n_0} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \partial \log Pr(H_i|\Phi)/\partial\lambda \\
&= -n_0 + \sum_{i=1}^{n-n_0} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})(\frac{c_i}{\lambda} - 1) = 0
\end{aligned} \tag{2.12}$$

The resulting closed form solution is given by:

$$\widehat{\lambda}^{(t+1)} = \frac{\sum_{i=1}^{n-n_0} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) c_i}{n} \tag{2.13}$$

2.2.3.2 Conditional Poisson Assumption

In general, datasets are comprised only of individuals with at least one detectable parasitic clone and it is assumed that uninfected individuals convey no information about haplotype frequencies. We now describe estimation under the conditional Poisson model in which we condition on at least one infection, as given in the following equation:

$$\phi_c^*(\lambda) = \begin{cases} (\lambda^c/c!)/(e^\lambda - 1) & c > 0 \\ 0 & c = 0 \end{cases} \tag{2.14}$$

14

The mean of this conditional Poisson distribution is given by:

$$\mu = \lambda/(1 - e^{-\lambda}) \tag{2.15}$$

Equation 2.1 for the haplotype combination probabilities is now replaced by:

$$Pr^*(H_i|\Phi) = \phi^*_{c_i}(\lambda)\frac{c_i!}{\delta_{i1}!\ldots\delta_{iK}!}\prod_{k=1}^{K}\theta_k^{\delta_{ik}} \tag{2.16}$$

Maximizing the conditional expectation of the complete data log likelihood results in the haplotype frequency estimates given by Equation 2.11 where now the summation is over all individuals.

The estimation of λ is achieved by solving the following equation.

$$
\begin{aligned}
\frac{\partial E\left[\log L(\Phi|\mathbf{X}_{com})\right]}{\partial\lambda} &= \sum_{i=1}^{n}\sum_{H_i\in\mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\partial\log Pr^*(H_i|\Phi)/\partial\lambda \\
&\propto \sum_{i=1}^{n}\sum_{H_i\in\mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})(\frac{c_i}{\lambda} - \frac{e^\lambda}{e^\lambda - 1}) = 0
\end{aligned}
\tag{2.17}
$$

This equation does not have a closed form solution and a Newton-Raphson procedure is employed. After we get the estimated λ, Equation 2.15 is used to get the estimate of the mean number of infections per infected individual.

2.2.4 Individual Predictions and Statistical Inference

Notably, the resulting parameter estimates can be used to infer individual level haplotypes and estimate the associated level of uncertainty. As described above, a set of haplotype combinations is consistent with each observed genotype. Based on the

final estimates of the haplotype frequences and the rate parameter, λ, the posterior probabilities of each haplotype set can be calculated. This allows us to determine the most probable haplotype set of a single individual, which may ultimately be relevant for making individual level treatment decisions.

Formal testing of population level parameters is also tenable. In order to test hypotheses involving haplotype frequencies, estimation of the corresponding variance-covariance matrix is needed. This estimate can be computed by inverting the observed information matrix. Within the E-M framework, this is computed via Louis' method [29]. An alternative approach is to approximate the observed information matrix with the empirical observed information matrix [32], given by:

$$I_e(\Phi; \mathbf{X}) = \sum_{i=1}^{n} s_i(\Phi) s_i^T(\Phi)\big|_{\Phi=\hat{\Phi}} \tag{2.18}$$

where $s_i(\Phi)$ is the score function from the observed data likelihood for the *ith* individual and can be computed as described in [31] and [25]:

$$s_i(\Phi) = E\{\partial[\log L_i(\Phi|X_i^{(com)})]/\partial\Phi|X_i^{(obs)}\} \tag{2.19}$$

2.3 A Simulation Study

A simulation study was conducted in order to characterize the performance of the maximum likelihood estimator under each of the 3 model assumptions for the number of infections: fixed, UP and CP. The values of C and λ were chosen to approximate the estimated values in the example provided in Section 2.4 and are presented in the rows of Table 2.1(a). Models considered include 2 and 3 loci with varying numbers of observed nucleotides at each, as described by the columns in Table 2.1(a). A

16

Table 2.1. Simulation Results

		Model 1[†]		Model 2[‡]		Model 3[*]	
		CR	CI-length	CR	CI-length	CR	CI-length
Fixed	C=2	(0.94, 0.96)	(0.023, 0.045)	(0.94, 0.96)	(0.036, 0.040)	(0.94,0.96)	(0.025,0.046)
	C=3	(0.95, 0.96)	(0.024, 0.040)	(0.94, 0.96)	(0.043, 0.045)	(0.94,0.96)	(0.036,0.058)
	C=6	(0.93, 0.95)	(0.035, 0.046)	(0.93, 0.95)	(0.084, 0.090)	(0.93,0.95)	(0.06,0.13)
CP	λ=0.8	(0.95, 0.96)	(0.031, 0.28)	(0.94, 0.96)	(0.055, 0.22)	(0.94,0.96)	(0.029,0.23)
	λ=2.0	(0.93, 0.95)	(0.027, 0.39)	(0.94, 0.96)	(0.047, 0.27)	(0.94,0.96)	(0.028,0.28)
	λ=3.0	(0.95, 0.97)	(0.029, 0.46)	(0.94, 0.96)	(0.048, 0.31)	(0.94,0.96)	(0.031,0.34)
UP	λ=1.5	(0.93, 0.96)	(0.027, 0.19)	(0.94, 0.95)	(0.049, 0.17)	(0.93,0.96)	(0.027,0.16)
	λ=2.0	(0.95, 0.96)	(0.026, 0.25)	(0.94, 0.96)	(0.045, 0.19)	(0.92,0.95)	(0.028,0.20)
	λ=3.0	(0.94, 0.96)	(0.027, 0.35)	(0.94, 0.96)	(0.045, 0.27)	(0.93,0.97)	(0.032,0.27)

(a) Overall results. The range of the coverage rates (CR) and the average length of the 95% confidence intervals (CI-length) are reported. [†] Model 1 has two alleles at each of two loci. The 4 haplotype probabilities are 0.64, 0.16, 0.16 and 0.04. [‡] Model 2 has two alleles at the first locus and three alleles at the second locus. The 6 haplotype probabilities are 1/6. [*] Model 3 has two alleles at each of three loci. The 8 haplotype probabilities are 0.24, 0.24, 0.16, 0.16, 0.06, 0.06, 0.04 and 0.04.

	$\lambda = 0.8$				$\lambda = 2.0$				$\lambda = 3.0$			
Parameter	Est	SE	CR	CI-length	Est	SE	CR	CI-length	Est	SE	CR	CI-length
$\theta_1 = 0.64$	0.64	0.02	0.95	0.078	0.64	0.015	0.95	0.06	0.64	0.014	0.96	0.056
$\theta_2 = 0.16$	0.16	0.014	0.95	0.053	0.16	0.011	0.93	0.043	0.16	0.011	0.95	0.042
$\theta_3 = 0.16$	0.16	0.014	0.95	0.056	0.16	0.011	0.95	0.043	0.16	0.011	0.97	0.042
$\theta_4 = 0.04$	0.04	0.008	0.96	0.031	0.04	0.007	0.95	0.027	0.04	0.007	0.96	0.029
λ	0.8	0.072	0.95	0.28	2	0.098	0.95	0.39	3	0.12	0.95	0.46

(b) Simulation Result for Model 1 Under CP Assumption. The mean estimate of the parameter (Est), the empirical standard error (SE), the cover rate (CR), and the average confidence interval length (CI-length) are shown.

more detailed summary of the simulation results for model 2 under CP assumption is presented in Table 2.1(b).

Under each model (fixed, UP and CP), $B = 500$ data sets with $n = 1000$ were generated. Using the algorithm appropriate for the underlying model, maximum likelihood estimates were obtained. The range of the Coverage rates (CR) and the average length of 95% confidence intervals (CI-length) are reported in Table 2.1 as measures of performance. CRs are calculated as the fraction of the 95% confidence

Table 2.2. Sample Genotype Data

Index	Genotypes of polymorphic region CSP-TH3											
1	T	G	A	A	C	G	C	C	G	A	G	A
2	T	G	A	A	C	G	C	C	G	A	G	A/C
3	T	G	A	A	C	G	C	C	G	A	G	C
4	T	G	A	A	C	G	C	C/G	G	A	G	A
5	T	G	A	A	C	G	C	C/G	G	A	G	A/C
6	T	G	A	A	C	G	C	G	G	A	G	A
7	T	G	A/G	A	C	G	C	C	G	A	A/G	A/T
8	T	G	G	G	T	A	C	G	G	A	G	A
9	T	A	A	A	C	G	C	C	G	A	G	C
10	T	G	A/G	A/G	C	G	C	C/G	G	A	G	A
11	T	G	A/G	A/G	C/T	G	C/G	C/G	A/G	A	G	A
12	T	G	A/G	A/G	C/T	A/G	C	G	G	A	G	A
13	T	G	A/G	A/G	C/T	G	C/G	C/G	A/G	A	G	A
⋮						...						

There are 12 loci in the csp-th3 region under consideration. Among them, 10 loci are polymorphic in our sample. A total of 55 unique genotypes are observed and 13 are presented above for illustration.

intervals that covered the true parameter value. For example, in model 1 with two alleles at each of two loci, there are four unknown parameters (haplotype probabilities) which were set to 0.64, 0.16, 0.16, and 0.04 under our simulation scenario. Under the CP assumption with $\lambda = 2$, as shown in Table 2.1(b), the CRs were 0.95, 0.93, 0.95, 0.95, and 0.95, for the 4 haplotype probabilities and λ, respectively. The average CI-lengths were 0.06, 0.043, 0.043, 0.027 and 0.39, respectively. Thus the range of CR and CI-length are reported in Table 2.1 as ranges equal to (0.93, 0.95) and (0.027, 0.39). These results indicate that the frequentist properties of our algorithms resulted in well calibrated interval estimates.

2.4 Example

In this section we present the results of applying the EM approach to data arising from a cross-sectional study of $n = 341$ malaria infected children from 3 African nations. Analysis is stratified by country in order to characterize potential regional differences. Data on 12 loci within one polymorphic region (csp-th3) within CSP are considered. A sample of observed genotype data is given in Table 2.2.

The CP assumption is most appropriate in this setting since the number of infections is variable and the data are comprised of samples with at least one detectable infection. In order to assess the sensitivity to the distributional assumption, however, we performed the analysis under both the CP and the UP assumptions. Under the CP assumption, the estimated population infection rates (λ) in Uganda, Cameroon and Sudan are 0.87, 1.08 and 0.75, respectively. The corresponding estimates of the mean numbers of infections per infected individual are $\mu = 1.50$, 1.64 and 1.42, respectively. Under the UP assumption, on the other hand, the estimated infection rates are 1.66, 1.80 and 1.60, respectively.

As expected, the estimated rates are higher for the UP approach since this approach assumes that uninfected individuals are included in the sampling design and the number of such individuals is equal to 0. In fact, using the λ from the CP approach, the expected number of uninfected individuals (n_0) in the sample of Uganda, Cameroon, and Sudan are 97.72, 75.63, and 52.62, respectively. If these numbers, instead of 0s, are used in the UP approach for n_0, then the results for UP and CP are similar (data not shown). Interestingly, in both models, the estimated number of infections is slightly higher in Cameroon than the other two countries. The estimated haplotype frequencies by country are similar for the two approaches.

Estimated haplotype frequencies by country under the CP assumption are presented in Table 2.3. This result suggests that while some haplotypes have similar frequencies across the three geographic regions (e.g. haplotype 1), there does appear to be variability across regions in the frequencies of other haplotypes. For example, haplotype 4 has an estimated frequency of about 5% in Uganda and Cameroon while its frequency in Sudan is estimated to be 22%.

The Wald test is used to test formally for regional haplotype differences. Specifically, we test the null hypotheses $H_0 : \theta_{i,j} = \theta_{i,j'}$, where θ is the frequency of haplotype i ($i = 1, \cdots, 16$) for country j ($j =$ Uganda, Cameroon, and Sudan). The test statis-

Table 2.3. Estimated Haplotype Frequencies by Region

Index	Haplotype	Uganda	Cameroon	Sudan
1	T G A A C G C C G A G C	0.34	0.39	0.33
2	T G A A C G C C G A G A	0.24	0.25	0.31
3	T G A A C G C G A A G A	0.10	0.07	–
4	T G A A C G C G G A G A	0.05	0.06	0.22
5	T G G G T A C G G A G A	0.05	0.03	0.02
6	T G G G C G C G G A G C	0.04	–	–
7	T G A A C G C C A A G A	0.04	0.06	–
8	T G G A C G C C G A G C	0.04	0.01	–
9	T G A A C G C G G A G C	0.03	0.07	–
10	T G G G C A C G G A G A	0.03	0.00	–
11	T G G G T G C G G A G A	0.01	–	0.06
12	T G G A C G C C G A A T	0.01	0.03	0.01
13	T G G G C G A G A A G A	0.01	–	–
14	T G G A C G C C G A G A	0.01	–	0.03
15	C G A A C G C G G G G A	–	–	0.01
16	T G G G T G C C G A G A	–	–	0.02

$N = 135$, 148 and 58 individuals were included in the analysis for Uganda, Cameroon and Sudan, respectively. Haplotypes with an estimated within country population frequency of at least 0.01 are reported.

tic corresponding to the null hypothesis $H_0 : \theta_{4,Sudan} = \theta_{4,Cameroon}$ is 3.5, which is significant at the 0.05 level after a Bonferroni correction is applied to adjust for 22 pairwise comparisons. Finally, posterior probability estimates of haplotype sets for each individual with the observed genotype are presented in Table 2.4. Again, there is reasonable consistency across the three countries though differences do appear to be present. Specifically for genotype 5, the probability that the true clones are single copies of haplotypes 4 and 1 is 65% in Sudan and only 31% in Cameroon. This is a reflection of the variable frequency estimates of haplotype 4 presented in Table 2.3.

2.5 Discussion

In this manuscript, we describe a novel model fitting approach to arriving at maximum likelihood estimates of haplotype frequencies in a population of children multiply infected with the parasite that causes malaria. Characterizing the genetic variability of the parasite, and particularly how polymorphisms align on a single clonal copy, will have broad implications for vaccine development efforts that target large genetic

Table 2.4. Estimated Posterior Probabilities for Each Haplotype Combination

Genotype (N^*)	Uganda		Cameroon		Sudan	
	Haplotype Set	p	Haplotype Set	p	Haplotype Set	p
1 (19,20,10)	2	0.90	2	0.87	2	0.89
	2 2	0.09	2 2	0.12	2 2	0.10
	2 2 2	−	2 2 2	0.01	2 2 2	0.01
2 (8,12,4)	1 2	0.77	1 2	0.70	1 2	0.78
	1 1 2	0.11	1 1 2	0.15	1 1 2	0.10
	1 2 2	0.08	1 2 2	0.09	1 2 2	0.09
3 (33,41,16)	1	0.86	1	0.80	1	0.88
	1 1	0.13	1 1	0.17	1 1	0.11
	1 1 1	0.01	1 1 1	0.02	1 1 1	−
4 (2,1,5)	4 2	0.88	4 2	0.85	4 2	0.82
	4 2 2	0.09	4 2 2	0.11	4 2 2	0.09
	4 4 2	0.02	4 4 2	0.03	4 4 2	0.07
5 (1,8,1)	4 1	0.46	4 1	0.31	4 1	0.65
	9 2	0.20	9 2	0.24	4 1 2	0.15
	4 1 2	0.10	9 1 2	0.10	4 1 1	0.08
6 (5,4,10)	4	0.98	4	0.97	4	0.92
	4 4	0.02	4 4	0.03	4 4	0.08
	4 4 4	−	4 4 4	−	4 4 4	−
7 (1,2,1)	12 2	0.89	12 2	0.86	12 2	0.87
	12 2 2	0.09	12 2 2	0.11	12 2 2	0.10
	12 2 2 2	−	12 12 2	0.02	12 14 2	0.02
8 (5,3,2)	5	0.98	5	0.98	5	0.99
	5 5	0.02	5 5	0.02	5 5	−
	5 5 5	−	5 5 5	−	5 5 5	−
. . .						

$N^* = (n_1, n_2, n_3)$ denotes the number of individuals within each country who present with the corresponding genotype. Genotypes are indexed in Table 2.2. Haplotypes are indexed in Table 2.3. − indicates a posterior probability estimate of less than 0.01. For comparison, only the genotypes present in all three countries are reported.

regions. Notably, the approach we describe for a fixed number of clones ($c > 2$) may be useful for polyploidy populations (e.g. goldfish, salmon, bread wheat, etc.), in which the number of chromosomes is more than 2 but the same across all units. Application of this approach to individuals multiply infected with HIV or carrying multiple viral mutations is also tenable.

Our approach offers two primary advantages over existing methods. First, the computational efficiency of our algorithm allows us to characterize a large number of sites. In the example described in Section 2.4, we illustrate straightforward implementation of this method in the context of 10 variable sites. This distinguishes our approach from the hill climbing method of [20]. Application of our method to the data for two variable loci described in [20] resulted in consistent estimates (results not shown.) Note that our approach also differs from [20] since we provide a testing method. Our method also allows for a variable number of clones within an individual, making it more flexible than approaches designed for diploid populations.

Several extensions that address the limitations of the proposed approach will provide further insight into the genetic variability and determinants of disease. Notably, it is assumed that the information on uninfected children is non-informative. That is, our analysis is based only on a cross-sectional study of children who were infected and does not consider the potentially informative data arising from children who could have been infected but were not. Application of causal inference methods may be appropriate in this setting. Further extensions would also allow us to relax the assumption that the observed number of heterozygous sites within an individual is not informative. Finally investigating methods for correlated data that provide a framework for evaluating changes over time would provide further insight into the molecular evolution of the parasite.

CHAPTER 3

ESTIMATING AND TESTING HAPLOTYPE-TRAIT ASSOCIATIONS IN NON-DIPLOID POPULATIONS

SUMMARY: Malaria is an infectious disease caused by a group of parasites of the genus *Plasmodium*. Characterizing the association between polymorphisms in the parasite genome and measured traits in an infected human host may provide insight into disease etiology while ultimately informing new strategies for improved treatment and prevention. This, however, presents an analytic challenge since individuals are often multiply infected with a variable and unknown number of genetically diverse parasitic strains. In addition, data on the alignment of nucleotides on a single chromosome, commonly referred to as haplotypic phase, is not generally observed. An expectation maximization algorithm for estimating and testing associations between haplotypes and quantitative traits has been described for diploid (human) populations. We extend this method to account for both the uncertainty in haplotypic phase and the variable and unknown number of infections in the malaria setting. Further extensions are described for the HIV/AIDS quasi-species setting. A simulation study is presented to characterize method performance. Application of this approach to data arising from a cross-sectional study of $n = 126$ multiply infected children in Uganda reveals some interesting associations requiring further investigation.

3.1 Introduction

Several methods and related extensions for characterizing population level haplotype frequencies and haplotype-trait associations in human populations have been

described [8, 40, 45, 39, 37, 25, 27, 11]. In the present manuscript, we propose a further extension of the EM approach for haplotype-trait association studies [25, 27] for infectious disease settings. Here interest lies similarly in characterizing the relationship between genetic information and a trait; however, in the infectious disease context, the genetic information is typically measured on the infectious agent (such as a parasite or virus) rather than the human. In both cases, we assume that the trait is a host (human) level measurement. To elucidate the challenges inherent to this setting, we begin by defining some terminology.

Consistent with the nomenclature adopted in several recent manuscripts, *genotype* is used to refer to the observed genetic information (nucleotides) at polymorphic sites. The term *haplotypic phase* refers to the alignment of nucleotides on a single homolog within a linked region of DNA. *Homologs* are double-stranded segments of DNA that contain information for the same biological or clinical trait but have potentially different genetic codes, commonly referred to as *alleles*. For example, humans, as diploid organisms, carry exactly two homologs (with the exception of the sex chromosomes), one inherited from each parent. In association studies of unrelated individuals, haplotypic phase is generally unobservable. That is, based on the observed genetic information, and specifically if an individual in heterozygous at two or more sites within a linked region of DNA, the alignment of nucleotides on each chormosomal copy is unknown.

In the context of infectious disease investigations, a parallel terminology is used. Here the term *strain* is used to refer to a parasite or viral lineage arising from a single ancestry. Application of population based sequencing techniques similarly renders the specific combination of nucleotides on a single strain unobservable. Notably, the methods cited above for the human genetics setting, were specifically developed to account for potential uncertainty in haplotypic phase, inherent in population-based investigations of unrelated individuals.

24

Our investigation is motivated by a study of the human pathogenic species *Plasmodium falciparum*, the group of parasites that cause malaria. Here interest lies in characterizing associations between genetic polymorphisms in the haploid parasite and clinical measures of disease severity, such as red blood cell count or the amount of parasite in plasma. In this setting, multiple infections can arise as a result of two or more singly infected mosquitoes taking blood meals from the same individual, an infected mosquito taking blood meals over several days, or a single multiply infected mosquito taking a blood meal from an individual. These three settings are indistinguishable from a data analytic perspective and all result in the presence of multiple strains within a single human host. In general, the observed genotype data consist of the set of bases present at each location of the genome across the entire population of organisms within a single host. Thus, as in the human genetics setting, the specific alignment of these bases on a single strain is generally unobservable. This constitutes the first analytic challenge.

The second challenge, rendering the infectious disease setting unique from human investigations, is that the *number* of infections, i.e. the *number* of strains, is unknown and this number can vary across individuals. This presents an additional analytic challenge and serves as the motivation for our present research. Consider, for example, an individual who is infected with up to 4 strains. In this case, between 1 and 4 bases will be observed at each site on the genome. Now suppose the observed genotype for this individual is A_1/A_2 for site 1 and B_1/B_2 for site 2. In this simple case, there are four possible haplotypes: $h_1 = (A_1, B_1), h_2 = (A_1, B_2), h_3 = (A_2, B_1)$ and $h_4 = (A_2, B_2)$. The precise combination of these haplotypes within this individual is not observable. In a human population, the number of homologs is fixed at 2 and therefore, the truth could be (h_1, h_4) or (h_2, h_3). However, in the malaria setting, since the number of strains within each person is also unobserved, the number of copies of each haplotypes is unknown. In this case, the true haplotype combination

could be (h_1, h_4), (h_2, h_3), (h_1, h_4, h_4) or (h_1, h_1, h_4, h_4), etc. and depends on whether the individual has 2, 3 or 4 infections. Note that two distinct strains may have the same haplotype for the gene under consideration and thus we include, for example, (h_1, h_4) and (h_1, h_4, h_4) as two distinct possibilities. we propose an EM-type algorithm that additionally takes into account information on a measured trait. This provides a comprehensive framework for simultaneous estimation of population haplotype frequencies and haplotype-trait associations. In previous work, we describe an expectation-maximization (EM)-type algorithm for estimating haplotype frequencies in the malaria setting that uses only the observed genotype data [26]. This prior work, while extending the methods of [8] and [20], does not take into account phenotypic or clinical information about the host. In the present manuscript, Thus the method presented represents an extension of [26] to incorporate trait information as well as an extension of [25] and [27] to the non-diploid setting.

An underlying premise motivating our research is that haplotypes may explain variability in a measured trait that is not fully captured by consideration of genotype data alone. In human genetic settings, haplotype-based investigations are important if the polymorphisms under consideration are in linkage disequilibrium with the true disease causing variant, but are not themselves causal. In the malaria settings, the specific combinations of nucleotides on a single strain may be relevant to protein production and ultimately, to parasite fitness. The method presented herein provides the framework for evaluating these potential associations.

In the following Section, we describe an extension of the EM framework for estimation and inference under several models for the distribution of the number of infections. In Section 3.3, this approach is applied in a simulation study as well as to data arising from a cohort of $n = 126$ multiply infected children from Uganda. Section 3.4 describes extensions for the HIV quasi-species setting in which multiple strains can arise from repeat infections though more generally, this is a result of ex-

ternal pressures, such as treatment exposures. Finally, in Section 3.5 we provide a discussion of our findings.

3.2 Methods

We begin in this section by outlining our notation and the structure of the data. We then describe three approaches to estimation of the effect of haplotypes on a quantitative trait that each involves different assumptions about the distribution of the number of infections: (1) First, we assume the number of infections within a host is fixed at a constant $C > 0$; (2) Second, we assume this number follows a conditional Poisson distribution where we condition on the presence of at least one infection; and (3) Third, we make no assumption about the distribution of the number of infections and estimate separately the probabilities of having exactly c infections where $c = 1, 2, \ldots, C$ for C sufficiently large. Finally, a formal testing procedure is described.

3.2.1 Notation

Let $\mathbf{G} = (G_1, \ldots, G_n)$ where G_i is the unphased (observed) multi-site genotype for individual i. Further suppose $\boldsymbol{\mathcal{H}} = (\mathcal{H}_1, \ldots, \mathcal{H}_n)$ where $\mathcal{H}_i$ represents the combination of haplotypes within individual i. In general, $\mathcal{H}_i$ is not known and multiple values of $\mathcal{H}_i$ are consistent with G_i. The set of all haplotype combinations that are consistent with G_i is denoted by $\mathcal{S}(G_i)$. Let $h_1, \ldots, h_K$ denote the K possible haplotypes over all observed individuals and define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ where θ_k is the population frequency of h_k. Now let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ where Y_i is the trait for $i = 1, \ldots, n$. We model $\mathbf{Y}$ using the generalized linear model (GLM) such that the expected value of Y_i is related to the linear predictor $\begin{bmatrix} \mathbf{X}_i^T & \mathbf{H}_i^T \end{bmatrix} \boldsymbol{\beta}$ through a link function g:

$$g(E[Y_i]) = \begin{bmatrix} \mathbf{X}_i^T & \mathbf{H}_i^T \end{bmatrix} \boldsymbol{\beta} \tag{3.1}$$

where $\mathbf{X}_i$ is a vector of environmental or demographic covariates, including the intercept as the first element, $\mathbf{H}_i$ is a vector of numerical codes for $\mathcal{H}_i$ and $\boldsymbol{\beta}$ is the corresponding parameter vector. For a quantitative trait, $g(\cdot)$ reduced to the identity link. Since the haplotype combination for individual i is potentially unobserved, we consider all possible $\mathcal{H}_i$ consistent with the observed genotype data, as described below in Section 3.2.2. Note $\mathbf{H}_i$ can take many forms depending on the specific genetic model. For example, we may define $\mathbf{H}_i$ as a $K \times 1$ vector of indicators for the presence of a specific dominant haplotype in individual i. Alternatively, we can set the kth element of $\mathbf{H}_i$ equal to the number of copies of h_k in individual i, corresponding to an additive genetic model. Further discussion of formulations for this design matrix are given in [27].

3.2.2 Estimation

In this section we describe the general EM framework for estimation, assuming a given distribution for the number of infections. We then elaborate on this algorithm for each of three distributional assumptions. First note that for the GLM framework, we assume that the probability density of $\mathbf{Y}$ is from an exponential family, given by:

$$Pr(\mathbf{Y}|\mathbf{X}, \mathbf{H}, \boldsymbol{\beta})$$

$$= L(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \mathbf{H}) = \prod_{i=1}^{n} exp \left\{ \frac{(Y_i \begin{bmatrix} \mathbf{X}_i^T & \mathbf{H}_i^T \end{bmatrix} \boldsymbol{\beta} - b \left(\begin{bmatrix} \mathbf{X}_i^T & \mathbf{H}_i^T \end{bmatrix} \boldsymbol{\beta} \right)}{a(\psi)} + c(Y_i, \psi) \right\}$$

$$(3.2)$$

where a, b, and c are known functions, ψ is a scale parameter and in our setting $\mathbf{H}$ is unknown. The ambiguity in $\mathbf{H}$ renders the haplotype-trait association study

28

a missing data problem and thus an EM-type algorithm is a natural choice for this setting. The EM algorithm, formalized by [5], involves first taking the conditional expectation of the complete data log likelihood (E-step), maximizing this with respect to the parameters of interest (M-step) and then iterating between these two steps until a convergence criterion is met. In our setting, the observed data consist of $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{G}$ and are denoted $\mathbf{X}^{(obs)}$, while the complete data consist of $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{G}$, and $\mathcal{H}$ and are denoted $\mathbf{X}^{(com)}$. Let Φ be the parameters of interest, as described in each of the following sections. The complete-data likelihood for Φ is thus given by:

$$L(\Phi|\mathbf{X}^{(com)}) = \prod_{i=1}^{n} Pr(Y_i|\mathbf{X}_i, \mathbf{H}_i, \boldsymbol{\beta}) Pr(\mathcal{H}_i|\boldsymbol{\theta}) \tag{3.3}$$

where $Pr(\mathcal{H}_i|\boldsymbol{\theta})$ is the corresponding haplotype set probabilities for the ith individual. Notably, this likelihood assumes the haplotype frequencies are independent of environmental/demographic information. In general, if departures from this assumption are tenable, a stratified analysis may be appropriate. As seen below, $Pr(\mathcal{H}_i|\boldsymbol{\theta})$ depends on the particular assumptions made with respect to the number of infections.

Let $\widehat{\Phi}^{(t)}$ be the estimate of Φ derived from the tth iteration of the EM algorithm. Formally, we have that the expectation of the complete data log likelihood conditional on the observed data and the current parameter estimates is given by:

$$
\begin{aligned}
& E\left[\log L(\Phi|\mathbf{X}^{(com)})|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right] \\
& = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \left[\log Pr(Y_i|\mathbf{X}_i, \mathbf{H}_i, \boldsymbol{\beta}) + \log Pr(\mathcal{H}_i|\boldsymbol{\theta})\right]
\end{aligned}
\tag{3.4}
$$

where:

29

$$p_{iH_i}(\widehat{\Phi}^{(t)}) = p(\mathcal{H}_i|\mathcal{H}_i \in \mathcal{S}(G_i), Y_i, \mathbf{X}_i, \widehat{\Phi}^{(t)}) = \frac{Pr(Y_i|\mathbf{X}_i, \mathbf{H}_i, \widehat{\boldsymbol{\beta}}^{(t)})Pr(\mathcal{H}_i|\widehat{\boldsymbol{\theta}}^{(t)})}{\sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} Pr(Y_i|X_i, H_i, \widehat{\boldsymbol{\beta}}^{(t)})Pr(\mathcal{H}_i|\widehat{\boldsymbol{\theta}}^{(t)})}$$
$$(3.5)$$

Next, we maximize the conditional expectation of the complete data log likelihood given in Equation 3.4. It is straightforward to show that the $(t+1)^{th}$ estimate of Φ can be obtained by finding the root for the following equations:

$$
\begin{aligned}
&\frac{\partial E\left[\log L(\Phi|\mathbf{X}^{(com)})|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right]}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \partial \log L(\boldsymbol{\beta}|Y_i, \mathbf{X}_i, \mathbf{H}_i)/\partial \boldsymbol{\beta} \\
&= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \frac{(Y_i - E(Y_i|\mathbf{X}_i, \mathbf{H}_i, \boldsymbol{\beta}))\left[\begin{array}{cc}\mathbf{X}_i^T & \mathbf{H}_i^T\end{array}\right]^T}{a(\psi)} \\
&= 0
\end{aligned}
$$
$$(3.6)$$

and

$$\frac{\partial E\left[\log L(\Phi|\mathbf{X}^{(com)})|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right]}{\partial \theta_k} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \partial \log Pr(\mathcal{H}_i|\boldsymbol{\theta})/\partial \theta_k = 0$$
$$(3.7)$$

As noted in [25] for the diploid setting, Equation 3.6 reveals that the regression parameter $\boldsymbol{\beta}$ can be estimated via weighted regression, where the weights are the

posterior probabilities of the haplotype sets for each individual, allowing us to use standard statistical software packages at this step. In the following subsections we describe estimation under specific assumptions for $Pr(\mathcal{H}_i|\boldsymbol{\theta})$. We assume algorithm convergence when $max\left(|\widehat{\Phi}^{(t)} - \widehat{\Phi}^{(t+1)}|/\widehat{\Phi}^{(t)}\right) < 1.0 \times 10^{-5}$. Alternatively, a convergence criterion can be based on the observed data likelihood, given by $\prod_{i=1}^{n}\sum_{H_i \in S(G_i)} Pr(Y_i|\mathbf{X}_i, \mathbf{H}_i, \boldsymbol{\beta})Pr(H_i|\theta)$.

3.2.2.1 Fixed number of infections

Let δ_{ik} denote the number of copies of haplotype h_k present in the haplotype combination $\mathcal{H}_i$. First suppose there are exactly C strains present in each individual where $C > 0$. That is, assume each individual has exactly C infections, where C is some known positive integer. Note that this implies $\sum_{k=1}^{K} \delta_{ik} = C$, where δ_{ik} ranges from 1 to C. $Pr(\mathcal{H}_i|\boldsymbol{\theta})$ of Equation 3.3 is thus given by:

$$Pr(\mathcal{H}_i|\boldsymbol{\theta}) = \frac{C!}{\delta_{i1}! \ldots \delta_{iK}!} \prod_{k=1}^{K} \theta_k^{\delta_{ik}} \tag{3.8}$$

In this case, $\Phi = (\boldsymbol{\beta}, \boldsymbol{\theta})$. Plugging Equation 3.8 into Equation 3.7, we have:

$$\frac{\partial E\left[\log L(\Phi|\mathbf{X}^{(com)})|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right]}{\partial \theta_k} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\left(\frac{\delta_{ik}}{\theta_k} - \frac{\delta_{iK}}{\theta_K}\right) = 0 \tag{3.9}$$

Resulting closed form solutions for $\hat{\theta}_k$ (see Appendix A) are given by:

$$\widehat{\theta}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{ik}}{nC} \tag{3.10}$$

3.2.2.2 Poisson assumption on the numbers of infections

In Section 3.2.2.1 we assume that the number of infections is fixed; however, in general this number may be variable for each individual. In this section we relax this assumption and instead assume a Poisson distribution on the number of infections per individual, as described in [20]. Since datasets are generally comprised only of individuals with at least one detectable infection, the conditional Poisson is considered. Let the Poisson model conditioning on at least one infection is given by:

$$
\phi_c(\lambda) = \begin{cases} (\lambda^c/c!)/(e^\lambda - 1) & c > 0 \\ 0 & c = 0 \end{cases}
\tag{3.11}
$$

where $\phi_c(\lambda)$ is the probability of having c infections. In this case, $\Phi = (\boldsymbol{\beta}, \boldsymbol{\theta}, \lambda)$. Since the number of strains c_i can be determined from $\mathcal{H}_i$, Equation 3.8 for the haplotype combination probabilities is now replaced by:

$$
Pr(\mathcal{H}_i|\boldsymbol{\theta}, \lambda) = Pr(\mathcal{H}_i, c_i|\boldsymbol{\theta}, \lambda) = \phi_{c_i}(\lambda)\frac{c_i!}{\delta_{i1}!\dots\delta_{iK}!}\prod_{k=1}^{K}\theta_k^{\delta_{ik}}
\tag{3.12}
$$

where c_i is the number of infections for i^{th} individual. Estimation of $\boldsymbol{\theta}$ proceeds similar to the setting in which C is fixed. Straightforward calculation (See Appendix B) leads to closed form solutions for $\hat{\theta}_k$ given by:

$$
\widehat{\theta}_k^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{ik}}{\sum_{i=1}^{n}\sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})c_i}
\tag{3.13}
$$

Estimation of λ is achieved by solving:

$$\frac{\partial E \left[\log L(\Phi | \mathbf{X}^{(com)}) | \mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)} \right]}{\partial \lambda} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \partial \log Pr(\mathcal{H}_i | \boldsymbol{\theta}, \lambda) / \partial \lambda$$

$$= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \left(\frac{c_i}{\lambda} - \frac{e^\lambda}{e^\lambda - 1} \right) = 0$$

$$(3.14)$$

There is no closed form for $\hat{\lambda}$ and a Newton-Raphson procedure can be employed. In this setting, the number of possible strains in an individual is not limited, which leads to an infinite sum in the E-step of the EM algorithm. In practice, we consider the number of strains to be limited by a large number (C) such that the probability of having more than C infections is small.

3.2.2.3 Semi-parametric approach

Finally, we consider the approach in which no assumptions are made about the distribution of the number of infections. In this approach, we estimate separately the probabilities of having exactly c infections where $c = 1, 2, \ldots, C$ for C sufficiently large. Let q_c be the probability of having c infections and define $\boldsymbol{q} = (q_1, ..., q_C)$, $\Phi = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{q})$. Equation 3.8 for the haplotype set probabilities is now replaced by:

$$Pr(\mathcal{H}_i | \boldsymbol{\theta}, \mathbf{q}) = \frac{c_i!}{\delta_{i1}! \ldots \delta_{iK}!} \prod_{k=1}^{K} \theta_k^{\delta_{ik}} \prod_{c=1}^{C} q_c^{I(c_i = c)} \qquad (3.15)$$

where $I(c_i = c)$ equals to 1 if $c_i = c$ and 0 otherwise. Estimation of $\boldsymbol{q}$ proceeds by solving:

$$\frac{\partial E\left[\log L(\Phi|\mathbf{X}^{(com)})|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right]}{\partial q_c} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \partial \log Pr(H_i|\boldsymbol{\theta}, \mathbf{q})/\partial q_c$$

$$= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \left[\frac{I(c_i = c)}{q_c} - \frac{I(c_i = C)}{q_C}\right] = 0$$

$$(3.16)$$

and resulting closed form solutions (See Appendix C) for $\hat{q}_c$ are given by:

$$\hat{q}_c^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})I(c_i = c) \qquad (3.17)$$

3.2.3 Inference

Wald tests are used to test hypotheses of haplotype-trait associations. In order to do this, estimates of the model parameters and the corresponding variance-covariance matrix are needed. Estimation of the variance-covariance matrix proceeds by inverting the observed information matrix, which is computed via Louis' method within the EM framework [29]. An alternative approach is to approximate the observed information matrix with the empirical observed information matrix which can be computed by [32]:

$$I_e(\Phi; \mathbf{X}) = \sum_{i=1}^{n} s_i(\Phi)s_i^T(\Phi)|_{\Phi=\hat{\Phi}} \qquad (3.18)$$

where $\hat{\Phi}$ is the estimates of the parameters in the last EM interation and $s_i(\Phi)$ is the score function from the observed data likelihood for the ith individual. The score is given by [31]:

$$s_i(\Phi) = E_\Phi\{\partial \log L_i(\Phi|X_i^{(com)})/\partial\Phi|X_i^{(obs)}, \widehat{\Phi}\} \tag{3.19}$$

For example, under the fixed assumption, we have:

$$s_i(\Phi) = \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}) \begin{bmatrix} [Y_i - E(Y_i|\mathbf{X}_i, \mathbf{H}_i, \boldsymbol{\beta})] \begin{bmatrix} \mathbf{X}_i^T & \mathbf{H}_i^T \end{bmatrix}^T /a(\psi) \\ (\delta_{i1}/\theta_1 - \delta_{iK}/\theta_K) \\ \vdots \\ (\delta_{ik-1}/\theta_{k-1} - \delta_{iK}/\theta_K) \end{bmatrix} \tag{3.20}$$

3.3   Data examples

In the following simulation study and real data example we focus on a quantitative trait for ease of presentation. In this case, $g(\cdot)$ of Equation 3.1 is set equal to the identity link and we have the following linear regression model:

$$Y_i = \begin{bmatrix} \mathbf{X}_i^T & \mathbf{H}_i^T \end{bmatrix} \boldsymbol{\beta} + \epsilon_i \tag{3.21}$$

We further assume the ϵ_i are independent and normally distributed with mean 0 and variance given by σ^2. Notably, this model assumes homoscedasticity and is therefore

applicable when the standard deviation of the trait is constant over the values of **X** and **H**. In the real data example provided below, we have no biological reason to believe that there is a violation of this assumption, though in general, evaluation of the appropriateness of the homoscedasticity assumption can be achieved through close examination of residual plots.

3.3.1 Simulation study

In order to evaluate the performance of the methods described in Section 3.2, we conduct a simulation study and report the type-1 error rates (ER) and power under each of the three models for the number of infections: fixed, Poisson, and semi-parametic. For each individual, the simulation starts by generating the number of infections c. Under the fixed model, the number of infections is set equal to a constant C. Under the Poisson assumption, c is generated randomly from a conditional Poisson distribution with assumed rate parameters $\lambda = 2$ and $\lambda = 3$. Finally, under the semi-parametric approach, we assume that the number of infections c ranges from 1 to 4 with corresponding probabilities $\mathbf{q} = (0.3, 0.3, 0.2, 0.2)$.

Next we simulate the haplotype combination for each individual based on the multinomial distribution. Four haplotypes, given by $h_1 = (A_1, B_1), h_2 = (A_1, B_2), h_3 = (A_2, B_1)$ and $h_4 = (A_2, B_2)$, with corresponding population frequencies of $\boldsymbol{\theta} = (0.25, 0.35, 0.20, 0.20)$, are assumed. The trait, Y is generated using random sampling with the error generated from a normal distribution. A single haplotype effect is assumed with an effect size ranging from 0.2 to 0.8. For simplicity of presentation, we let $\sigma^2 = 1$ and vary β. In addition, we consider a model in which there is no haplotype effect, in which case the response is generated simply from a normal distribution with mean and variance equal to 1. In all cases, a dominant genetic model is assumed. For each configuration, $B = 200$ datasets with sample sizes of $n = 500$ are generated. Analysis is performed using genotype data and trait information only.

That is, we assume haplotypic phase and the number of infections is unknown and apply the methods described in Section 3.2 above.

Simulation results are provided in Table 3.1. Bias, coverage rates, power and ER are reported. Bias is defined as the absolute difference between the mean parameter estimates over the simulations and the true value. The estimated standard error of the parameter estimates based on the simulations is given by $\hat{se}$. The parameter β_1, the haplotype effect for the first haplotype $h_1 = (A_1, B_1)$, is varied across the simulations. Power is defined as the proportion of simulations in which we detect the true haplotype effect. The ER is the proportion of simulations for which an incorrect haplotype is detected, averaged over the haplotypes that are assumed to have no effect.

Under each of the three model assumptions and a range of haplotype effect sizes, the bias ranges from < 0.001 to 0.086 and the coverage rates are between 0.92 and 0.97. This suggests that our algorithms results in reasonably well calibrated interval estimates. As expected, the power for detecting the haplotype effect increases as the effect size increases from 0.0 to 0.8. In general, for samples of size of $n = 500$, we achieve $> 80\%$ power to detect moderate effect sizes of > 0.40. Notably, however, we see a reduction in power and an increase in the bias for β_1 as the number of infections (parasite strains) is increased from 2 to 4 under the fixed assumption. This is likely to be the result of increased ambiguity associated with more possible haplotype combinations within an individual as the number of infections (C) increases.

In order to evaluate the performance of the proposed method when the number of infections violates model assumptions, we conduct several sensitivity analyses. First, we perform estimation using the fixed approach, assuming the number of infections is equal to 2, when in fact the probabilities of having c infections for $c = 1 \ldots 5$ are all equal to 0.2. The results are presented in Table 3.2(a). Comparing this to correct application of the semi-parametric method (Table 3.1), we see a dramatic

power loss and a less severe, but noteworthy, decrease in coverage rates for both β and θ. In addition, the type-1 ER is substantially larger for $\beta_1 \geq 0.4$. Secondly, we perform estimation using the fixed approach, again assuming the number of infections is equal to 2, when in fact the number of infections arises from a conditional Poisson distribution with $\lambda = 2$. The results are presented in Table 3.2(b). Comparing these results to correct application of the Poisson approach with $\lambda = 2$ (Table 3.1), we see a more dramatic decrease in coverage rates for both β and θ. In addition, a significant decrease in power and increase in the type-1 ER are observed for $\beta \geq 0.2$. These findings support the use of the more sophisticated modeling approaches in these setting.

Next, we perform estimation using the Poisson approach when in fact the probabilities of having c infections for $c = 1 \ldots 5$ are all equal to 0.2 and present the results in Table 3.2(c). Here the modeling approach provides estimates of λ and from this, we calculate $\widehat{q_c}$ as $(\hat{\lambda}^c/c!)/(e^{\hat{\lambda}} - 1)$. As expected under this type of model misspecification, the coverage rates for q_c are very low $(0.12 - 0.15)$. Interestingly, the coverage rates for both β and θ remain at approximately 95% and the power and ER are reasonable, though slightly worse than using the correct model (Table 3.1). Finally, we evaluate performance in applying the semi-parametric approach when the number of infections actually arises from a Poisson with $\lambda = 2$. These results are given in Table 3.2(d) and as expected, we see a slight loss in power for the smaller effect sizes. For example, for an effect size of 0.4, the power of correctly using the Poisson approach is 0.87 (Table 3.1). Power for the semi-parametric approach is estimated to be 0.81. Since we are not incorporating knowledge about the distribution of the number of infections the loss in power is expected.

3.3.2 Multiply infected children with Malaria

Malaria is an infectious disease affecting millions of individuals globally. In fact, each year an estimated 1-3 million people die as a result of infection with the human pathogenic *Plasmodium* species, the group of parasites that causes malaria [3]. The majority of these deaths are in children under the age of 5 and in resource-constrained settings since current treatment options are costly or unavailable [17, 18]. Recent advances in sequencing technologies provide new opportunities for population-based genetic association studies to uncover complex relationships among genetic polymorphisms and measures of disease progression. Ultimately, these discoveries may help to inform novel strategies for vaccine development.

One of the biggest challenges in characterizing genotype-trait associations in this setting arises from the fact that individuals can be infected simultaneously with multiple parasitic strains. In the present investigation, we apply a novel approach to this challenge (see Section 3.2) to data arising from a cross-sectional study of $n = 126$ malaria infected children from Uganda. We focus on haplotypes in one polymorphic region (CSP-TH3) of the gene that encodes for a cellular adhesion domain of the circumsporozoite protein (CSP). CSP facilitates adhesion of the parasite to liver cells, a critical initial step in its replication process in a human host [44, 21]. The goal of our analysis is to uncover haplotype associations with red blood cell (RBC) count (log-transformed). RBC count is a well-known diagnostic tool for detecting anemia, a common and often lethal manifestation of malaria.

Data on 12 sites, 10 of which are polymorphic in our sample, are considered. Across all individuals, we see up to 3 different nucleotides at a site and within a single individual, between 1 and 2 nucleotides are present at any given site. A total of 35 unique genotypes are observed in our data and a sample of the data is provided in [26]. For computational purposes, the set of possible haplotypes is limited to those with estimated frequencies of greater than 0.01 where frequency estimates are

obtained using the approach of [26]. We assume a Poisson distribution and apply the approach of Section 3.2.2.2. A dominant genetic model is assumed, as in the simulation study.

Estimated haplotype effects on RBC and corresponding p-values for tests of the null hypotheses that these effects equal 0, are provided in Table 3.3. P-values are unadjusted for multiple comparisons. Using a Bonferroni adjustment, p-values less than $0.05/14 = 0.0036$ are considered significant at the 0.05-level. A significant association is observed between red blood cell count and the 3 haplotypes numbered 8, 11 and 12. Interestingly, the effect of carrying at least one copy of haplotype 11 appears to increase RBC count $e^{0.344} = 1.41$-fold, suggesting a potential protective effect. On the other hand, haplotypes 8 and 12 result in a lower RBC count with estimated decreases of $e^{-0.484} = 0.616$-fold and $e^{-0.137} = 0.872$-fold, respectively. Notably, the estimated number of individuals with each of these haplotypes (given by $126 * \widehat{\theta}_k$) is small and further confirmatory research is required to make firm conclusions.

3.4 Further extensions for the quasi-species setting

In the methods described above for estimation of haplotype effects on a trait, we incorporate population level haplotype frequencies. These frequencies can be thought of as the amount of each parasite strain circulating in the mosquito population that infects humans. Importantly, we assume that the frequencies within individuals reflect these population level parameters. In other words, the probability of being infected with a given strain does not depend on prior infections and is equal to the proportion of this strain in the general population. Patients infected with the human immunodeficiency virus (HIV) similarly host a population of viruses; however, the presence of such a quasi-species generally results from external pressures, such as drug exposures, rather than multiple repeat infections. As a result, the frequencies

of each haplotype within an individual may not reflect the true population level frequencies. This is evidenced, for example, by the existence of latent reservoirs of resistant variants that rapidly emerge in the presence of drug.

For this reason, rather than using population level haplotype frequencies in the HIV setting, we consider the probabilities that an individual in the target population carries a given haplotype. Note that while this distinction is subtle, it does require modification of the estimation approach described in Section 3.2. Again let G_i be the unphased (observed) multi-site genotype for the ith individual where $i = 1, \ldots, n$. Further suppose $\mathcal{H}_i$ represents the combination of *unique* haplotypes within individual i where $\mathcal{H}_i$ is generally unobservable and multiple values of $\mathcal{H}_i$ are consistent with G_i. We emphasize unique here since in the previously described approach, such a minimal set was not required. That is, we are now interested in whether an individual carries a specific haplotype and not in the number of copies. Again, the set of all combinations that are consistent with G_i is denoted $\mathcal{S}(G_i)$ and $h_1, \ldots, h_K$ denotes the K possible haplotypes over all observed individuals. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ where α_k is the probability that an individual carries at least one copy of h_k and define:

$$\delta_{ik} = \begin{cases} 1 & \text{if } h_k \text{ is present in } i^{th} \text{ individual} \\ 0 & \text{if } h_k \text{ is not present in } i^{th} \text{ individual} \end{cases} \tag{3.22}$$

Under the model given in Equation 3.1, the complete likelihood function can again be written as in Equation 3.3 where $Pr(\mathcal{H}_i | \boldsymbol{\theta})$ is replaced with:

$$Pr(\mathcal{H}_i | \boldsymbol{\alpha}) = \prod_{k=1}^{K} \alpha_k^{\delta_{ik}} (1 - \alpha_k)^{1 - \delta_{ik}} \tag{3.23}$$

In this case, estimation of the regression parameter β proceeds as described above and an estimate of α is obtained by finding the root of the following equation:

41

$$\frac{\partial E\left[\log L(\Phi)|\mathbf{X}^{(com)}|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right]}{\partial \alpha_k} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{i\mathcal{H}_i}(\widehat{\Phi}^{(t)}) \partial \log Pr(\mathcal{H}_i|\boldsymbol{\alpha})/\partial \alpha_k$$

$$= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{i\mathcal{H}_i}(\widehat{\Phi}^{(t)})(\frac{\delta_{ik}}{\alpha_k} - \frac{1 - \delta_{ik}}{1 - \alpha_k}) = 0 \tag{3.24}$$

Resulting closed form solutions (See Appendix D) for $\hat{\alpha}_k$ are given by:

$$\widehat{\alpha}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{i\mathcal{H}_i}(\widehat{\Phi}^{(t)})\delta_{ik}}{n} \tag{3.25}$$

3.5 Discussion

In this manuscript, we describe an approach to estimate and test haplotype-trait associations among individuals with multiple strains of an infectious agent. Three approaches to modeling the number of infections are described in Section 3.2 above. The first, which involves fixing the number of infections to be a constant c, is presented since it represents a natural extension of the diploid setting, within which $c = 2$ and our approach reduces to the EM method of [25]. Since in the infectious disease setting the number of infections is rarely known with certainty, this first approach may be more relevant to investigations of polyploidy organisms in which the number of homologous chromosomes is greater than 2, such as flatworms, goldfish, salmon and a variety of ferns and flowering plants. Note that the assumption of independent segregation made in Equation 3.8 needs to be addressed specifically for each of these settings.

Our simulation study suggests that application of the Poisson approach when in fact the numbers of infections are $c = 1, \ldots, 5$ with equal probabilities, results

in reasonable power and type-1 error rates but substantial bias in these probability estimates. The semi-parametric approach performs reasonably well under the Poisson model with a slight loss of power. Incorrect application of the fixed approach leads to more substantial losses in power, reductions in coverage rates and increases in type-1 error rates. Applications of the correct models lead to reasonable power and control of type-1 error rates.

Coupled with this investigation is the need for appropriate methods for controlling type-1 error rates in the context of multiple comparisons. In Section 3.3.2, we applied a Bonferroni correction to assess significance. Alternative single-step and step-down methods based on the false discovery rate and that account for the correlated nature of these tests [1, 2, 41] are also tenable. In addition, further consideration of resampling-based approaches and related extensions [43, 36, 10] may be appropriate. Extensions of the mixed-effects modeling approaches developed originally for the diploid setting [12, 11] would offer a single degree of freedom omnibus test for association across all haplotypes.

Notably, our analysis is limited to data arising from individuals who visited one of the designated clinics. This may lead to ascertainment bias for several reasons, including that the individuals under study exhibited symptoms severe enough to warrant at least one visit to the doctor. This is a potential limitation of the method described herein. Specifically, a population-level prevalence greater than 0 of infection by a strain that results in mild symptoms may result in overestimation of the frequencies of haplotypes that lead to more severe symptoms.

Application of this EM approach to a small cohort of children in Uganda revealed 3 potentially informative haplotypes within the CSP region of the parasite genome. In general, characterizing the association between polymorphisms in the parasite genome and measured traits in an infected human host may provide greater insight into disease etiology and help to inform new strategies for treatment and

vaccine development efforts. Drawing meaningful biological and clinical conclusions, however, will require further analysis. Specifically consideration of host level factors, such as host genetic profile and clinical or demographic features may be warrented. The methods described herein provide a general framework and the analytic tools to investigate such associations under several models of association and models for the numbers of infections.

Table 3.1. Simulation Results for dominant model under 3 assumptions

	β_1^*	Bias($\hat{se}$)† $\hat{\beta}_1$	–	$\bar{\hat{\boldsymbol{\theta}}}$	Coverage rates‡ β_1	–	$\bar{\boldsymbol{\theta}}$	Power**	ER***
FIXED									
	0.0	0.0038 (0.132)	–	0.0008 (0.016)	0.95	–	0.95	0.05	0.06
	0.2	0.0009 (0.138)	–	0.0005 (0.015)	0.96	–	0.95	0.35	0.07
C=2	0.4	0.0060 (0.138)	–	0.0013 (0.015)	0.96	–	0.95	0.82	0.06
	0.6	0.0002 (0.126)	–	0.0003 (0.016)	0.95	–	0.95	0.99	0.06
	0.8	0.0016 (0.122)	–	0.0008 (0.015)	0.94	–	0.95	1.00	0.05
	0.0	0.0035 (0.180)	–	0.0007 (0.018)	0.94	–	0.94	0.08	0.07
	0.2	0.0122 (0.181)	–	0.0009 (0.017)	0.95	–	0.95	0.22	0.08
C=3	0.4	0.0136 (0.187)	–	0.0006 (0.017)	0.95	–	0.95	0.59	0.08
	0.6	0.0265 (0.181)	–	0.0011 (0.017)	0.95	–	0.95	0.88	0.08
	0.8	0.0291 (0.177)	–	0.0004 (0.017)	0.95	–	0.94	0.97	0.07
	0.0	0.0128 (0.206)	–	0.0066 (0.019)	0.94	–	0.92	0.07	0.06
	0.2	0.0078 (0.223)	–	0.0037 (0.019)	0.97	–	0.94	0.20	0.09
C=4	0.4	0.0443 (0.212)	–	0.0065 (0.020)	0.96	–	0.94	0.38	0.06
	0.6	0.0856 (0.185)	–	0.0048 (0.020)	0.93	–	0.95	0.62	0.07
	0.8	0.0627 (0.197)	–	0.0046 (0.018)	0.92	–	0.95	0.88	0.06
POISSON		$\hat{\beta}_1$	$\hat{\lambda}$	$\bar{\hat{\boldsymbol{\theta}}}$	β_1	λ	$\bar{\boldsymbol{\theta}}$	Power**	ER***
	0.0	0.0098 (0.126)	0.0022 (0.111)	0.0025 (0.020)	0.96	0.94	0.94	0.04	0.05
	0.2	0.0011 (0.150)	0.0093 (0.105)	0.0011 (0.019)	0.95	0.95	0.95	0.41	0.05
λ=2	0.4	0.0001 (0.128)	0.0101 (0.089)	0.0013 (0.020)	0.96	0.96	0.97	0.87	0.06
	0.6	0.0240 (0.129)	0.0042 (0.116)	0.0018 (0.020)	0.94	0.98	0.96	1.00	0.03
	0.8	0.0160 (0.146)	0.0091 (0.104)	0.0012 (0.019)	0.96	0.95	0.94	0.99	0.05
	0.0	0.0022 (0.131)	0.0087 (0.123)	0.0017 (0.019)	0.96	0.97	0.94	0.04	0.03
	0.2	0.0312 (0.129)	0.0372 (0.124)	0.0027 (0.019)	0.95	0.96	0.95	0.44	0.04
λ=3	0.4	0.0002 (0.122)	0.0043 (0.137)	0.0017 (0.020)	0.94	0.96	0.95	0.91	0.05
	0.4	0.0055 (0.129)	0.0216 (0.137)	0.0009 (0.018)	0.93	0.96	0.94	0.99	0.06
	0.8	0.0120 (0.116)	0.0067 (0.126)	0.0024 (0.020)	0.97	0.96	0.94	1.00	0.06
SEMI-PARAMETRIC		$\hat{\beta}_1$	$\bar{\hat{q}}$	$\bar{\hat{\boldsymbol{\theta}}}$	β_1	$\bar{q}$	$\bar{\boldsymbol{\theta}}$	Power**	ER***
	0.0	0.0034 (0.117)	0.0112 (0.033)	0.0024 (0.019)	0.95	0.79	0.96	0.05	0.03
	0.2	0.0082 (0.108)	0.0119 (0.030)	0.0027 (0.018)	0.94	0.85	0.95	0.38	0.06
	0.4	0.0024 (0.118)	0.0119 (0.029)	0.0018 (0.018)	0.96	0.81	0.96	0.94	0.06
	0.6	0.0321 (0.141)	0.0132 (0.032)	0.0027 (0.019)	0.97	0.83	0.96	1.00	0.04
	0.8	0.0015 (0.116)	0.0119 (0.032)	0.0007 (0.018)	0.96	0.83	0.95	1.00	0.05

*β_1 is the effect of haplotype $h_1 = (A_1, B_1)$ on Y. †Bias is defined as the absolute difference between the mean of the estimate over the simulations and the true parameter value. ‡Coverage rate is defined as the proportion of simulations for which the true parameter value is within the corresponding 95% confidence interval. **Power is the specific power for the haplotype effect of the first haplotype h_1. ***ER is the type I error rate. $\bar{\hat{\boldsymbol{\theta}}}$ and $\bar{\hat{q}}$ denote averaging across all $\hat{\theta}$s and $\hat{q}$s, respectively. $\bar{\boldsymbol{\theta}}$ and $\bar{q}$ denote averaging across all θs and qs, respectively.

Table 3.2. Sensitivity Analysis to model mispecification

β_1^*	BIAS†		COVERAGE RATES‡		Power**	ER***
	$\hat{\beta}_1$	$\bar{\bar{\theta}}$	β_1	$\bar{\theta}$		
0.0	0.0016 (0.133)	0.0332 (0.044)	0.95	0.90	0.03	0.04
0.2	0.0441 (0.165)	0.0334 (0.045)	0.93	0.92	0.22	0.04
0.4	0.0810 (0.187)	0.0366 (0.042)	0.92	0.86	0.59	0.12
0.6	0.0761 (0.251)	0.0303 (0.041)	0.92	0.88	0.88	0.22
0.8	0.1081 (0.329)	0.0214 (0.044)	0.93	0.93	0.95	0.30

(a) Incorrect application of the fixed approach under semi-parametric data. The data are simulated assuming between 1 and 5 infections with equal probabilities of 0.20 while the estimation approach assumes $c = 2$ fixed infections. See Figure 1 legend for definitions of terms.

β_1^*	BIAS†		COVERAGE RATES‡		Power**	ER***
	$\hat{\beta}_1$	$\bar{\bar{\theta}}$	β_1	$\bar{\theta}$		
0.0	0.0158 (0.178)	0.0640 (0.104)	0.93	0.99	0.08	0.07
0.2	0.1112 (0.175)	0.0850 (0.083)	0.89	0.92	0.13	0.09
0.4	0.1499 (0.187)	0.0985 (0.065)	0.91	0.64	0.30	0.16
0.6	0.2177 (0.219)	0.0972 (0.068)	0.86	0.68	0.65	0.25
0.8	0.3546 (0.353)	0.0722 (0.092)	0.87	0.98	0.83	0.40

(b) Incorrect application of the fixed approach under Poisson distributed data. The data are simulated assuming a conditional Poisson distribution with $\lambda = 2$, while the estimation procedure assumes $c = 2$ fixed infections.

β_1^*	BIAS†			COVERAGE RATES‡			Power**	ER***
	$\hat{\beta}_1$	$\bar{\hat{q}}$	$\bar{\bar{\theta}}$	β_1	$\bar{q}$	$\bar{\theta}$		
0.0	0.0086 (0.115)	0.0492 (0.009)	0.0023 (0.022)	0.97	0.15	0.95	0.02	0.04
0.2	0.0110 (0.142)	0.0491 (0.009)	0.0019 (0.022)	0.95	0.14	0.95	0.37	0.07
0.4	0.0026 (0.129)	0.0489 (0.008)	0.0011 (0.020)	0.96	0.12	0.94	0.90	0.05
0.6	0.0039 (0.141)	0.0492 (0.008)	0.0010 (0.021)	0.94	0.13	0.96	0.99	0.05
0.8	0.0134 (0.102)	0.0492 (0.009)	0.0010 (0.020)	0.95	0.15	0.94	1.00	0.06

(c) Incorrect application of the conditional Poisson model. The data are simulated assuming between 1 and 5 infections with equal probabilities of 0.20.

β_1^*	BIAS†		COVERAGE RATES‡		Power**	ER***
	$\hat{\beta}_1$	$\bar{\bar{\theta}}$	β_1	$\bar{\theta}$		
0.0	0.0113 (0.114)	0.0027 (0.019)	0.96	0.95	0.04	0.05
0.2	0.0166 (0.123)	0.0025 (0.021)	0.95	0.95	0.34	0.04
0.4	0.0316 (0.147)	0.0025 (0.020)	0.97	0.96	0.81	0.04
0.6	0.0191 (0.115)	0.0022 (0.021)	0.95	0.95	1.00	0.05
0.8	0.0233 (0.121)	0.0010 (0.019)	0.94	0.94	1.00	0.04

(d) Incorrect application of the semi-parametric approach under Poisson distributed data. The data are simulated assuming a conditional Poisson distribution with $\lambda = 2$. The number of infections is assumed to range from 1 to 10.

Table 3.3. Estimated Haplotype Effects for Uganda

	Unique Haplotype	Est Freq ($\widehat{\theta}$)	Est Effect ($\widehat{\beta}$)	SE	P-value
1	T G A A C G C C G A G C	0.328	-0.108	0.099	0.278
2	T G A A C G C C G A G A	0.241	-0.066	0.092	0.471
3	T G A A C G C G A A G A	0.103	-0.032	0.106	0.762
4	T G A A C G C G G A G A	0.057	-0.148	0.150	0.324
5	T G G G T A C G G A G A	0.044	-0.257	0.151	0.089
6	T G G G C G C G G A G C	0.046	-0.081	0.240	0.737
7	T G A A C G C C A A G A	0.046	-0.023	0.165	0.891
8	T G G A C G C C G A G C	0.041	-0.484	0.133	<0.001*
9	T G A A C G C G G A G C	0.034	0.200	0.583	0.731
10	T G G G C A C G G A G A	0.022	0.159	0.331	0.631
11	T G G G T G C G G A G A	0.011	0.344	0.008	<0.001*
12	T G G A C G C C G A A T	0.005	-0.137	0.000	<0.001*
13	T G G G C G A G A A G A	0.011	0.292	0.806	0.717
14	T G G A C G C C G A G A	0.009	0.206	2.031	0.919

*Indicates haplotype effect on RBC is significantly different than 0 after applying a Bonferroni adjustment for multiple-comparisons. Results are based on a sample of size $n = 126$ and assume a Poisson model for the number of strains per individual.

CHAPTER 4

BAYESIAN MODELING WITH AMBIGUOUS CLUSTER IDENTIFIERS

SUMMARY: Mixed modeling is a useful approach for Characterizing haplotype-trait associations in the context of population-based association studies of unrelated individuals. In this setting, clusters are defined as groups of genetically similar individuals, for example, the individuals who carry a common pair of haplotypes. This presents an analytical challenge, however, since haplotypic phase (i.e. the alignment of bases on a single DNA strand) is generally unobservable. Therefore, the cluster identifier is ambiguous. In this paper, we describe a Bayesian method for estimation in this missing data setting. Two prior distributions for cluster effects are assumed. A simulation study is also presented to characterize method performance and assess sensitivity to distributional assumptions.

4.1   Introduction

Characterizing haplotype-trait associations in the context of population-based association studies of unrelated individuals presents several analytical challenges arising from: (1) the unobservable nature of haplotypic phase and (2) the large number of potentially informative haplotypes under study. Haplotypic phase refers to the alignment of alleles on a single homolog, inherited from a single parental genome, and is relevant in the context of studying genotype-trait associations if the genetic polymorphisms under investigation are markers for the true disease causing allele. Phase information is typically not observed in populations of unrelated individuals since ambiguity arises when heterozygosity is present at more than one locus within

48

a gene. In addition to the challenge of phase ambiguity, methods for characterizing haplotype-trait association require consideration of multiple haplotypes within a gene. Several methods and related extensions for characterizing haplotype-trait associations in human populations have been described [25, 27]. These methods estimate the haplotype-trait association in a generalized linear model frame work and work well when the number of haplotypes is small. The number of haplotypes, however, can be large and is an increasing function of the observed number of single nucleotide polymorphisms (SNPs), given by $M = 2^S$ where S is the number of biallelic SNPs under study within the corresponding gene. In that case, the number of coefficients will be large and estimation of halplotype effects will not be feasible.

In a recent manuscript, [11] describe the application of a mixed effects modeling framework to the setting where clusters are defined as groups of genetically similar individuals. This is a natural extension of the analysis of data arising from family-based studies in which clusters are defined as self-declared family units [24]. An expectation conditional maximization (ECM) approach is developed to account for uncertainty in the cluster identifiers arising from unobserved haplotypic phase. The primary advantage of the mixed model is that it addresses the well-known degrees of freedom problem that arises in the application of an analysis of variance (ANOVA) to this setting. The motivation for clustering based on pairs of haplotypes is multi-faceted. On the one hand, it is a natural grouping that is similar to grouping by family units since individuals within the same family tend to have similar genetic profiles. Secondly, this approach allows for simultaneous consideration of several polymophisms within a gene since haplotypes are defined based on multiple single nucleotide polymorphisms (SNPs). This is particularly relevant in the presence of statistical interaction among SNPs. Thirdly, as discussed above, haplotypes can capture more variability in a disease trait than genotype data alone, since SNPs are often markers for the true disease causing variant. Finally, consideration of pairs of haplotypes provides for discovery of genetic interaction between homologous chromosomes. In this manuscript, we describe two Bayesian approaches to estimation and testing in the context of mixed modeling with missing cluster identifiers. Importantly, we distinguish here between

latent cluster effects on the trait and latent group identifiers. In our context, both the cluster effects and the cluster identifiers are potentially unobservable. The approach we present is a natural extension of the Gibbs sampler described for mixed models and has a marked computational advantage over the ECM approach.

Two assumptions for random cluster effects are used in the proposed model. We first consider a single normal prior on the random effects and then describe a mixture modeling approach as a related method for discovering haplotype-trait associations. In the later, we relax the strict normality assumption described in previous work for the haplotype setting and assume random cluster effects arise from a discrete mixture distribution with a Dirichlet process prior. Methods for Bayesian mixture model fitting are well-described [6, 7, 23]. Interestingly, the Bayesian mixture modeling naturally handles the ambiguity in cluster identifiers through simply assigning each ambiguous individual to a single cluster. It also provides a powerful tool for discovery in the context of a large number of markers. On the other hand, this approach does not involve reconstructing individual level haplotypes. Thus further extensions are needed to make conclusions about the specific haplotypes contributing to variability in the trait and to estimate population level haplotype frequencies.

Bayesian methods for the analysis of data arising from population-based association studies of unrelated individuals have been described previously. For example, the pivotal work of [40] includes application of a Gibbs sampler to reconstruct individual level haplotypes while drawing strength from a population genetic coalesence model. Notably, this approach does not address haplotype-trait associations, which is the focus of the present manuscript. More recently, [38] apply a Bayesian latent class analysis for whole genome wide association data which, similar to the approach described herein, considers random effects that arise from a mixture distribution. Our approach, however, differs in that we consider haplotype data that are themselves unobservable, yielding a doubly latent class structure.

In the following section we describe two Bayesian approaches to fitting the linear mixed model in the context of unobservable cluster identifiers. In Section 4.3, we

characterize these methods through a simulation study. Finally, in Section 4.4 we offer a discussion of our findings.

4.2 Methods

In this section we present two Bayesian approaches to fitting a linear mixed model in the context of unobservable cluster identifiers. While these approaches are broadly relevant to settings with cluster ambiguity, we describe our methods in the context of the analysis of population-based genetic association data. In our setting, the aim is to characterize the association between haplotypes and a quantitative trait. We begin by defining the associated model and describing clusters based on haplotypes.

4.2.1 Mixed model for haplotype-trait associations

Let $\mathbf{y} = (y_1, \ldots, y_n)$ where y_i is the observed response for individual i and let $j = (1, \ldots, J)$index the J possible clusters. Let $\mathbf{c}_i = (c_{i1}, \ldots, c_{iJ})$ where $c_{ij} = 1$ if individual i is in cluster j. Throughout this manuscript, we assume the number of individuals in each cluster follows a multinomial distribution, given by:

$$\left[\sum_{i=1}^n c_i | \boldsymbol{\pi} \right] \propto \prod_{j=1}^J \pi_j^{\sum_{i=1}^n c_{ij}} \tag{4.1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ and π_j is the population probability of an individual belonging to the jth cluster

In general, because of the ambiguity of cluster identifier, $\mathbf{c}_i$ is latent. Let $\mathcal{S}(G_i)$ be the set of possible cluster membership for individual i. If $\mathcal{S}(G_i)$ has one element, then $\mathbf{c}_i$ is known. Let $\mathbf{b} = (b_1, \ldots, b_J)$ denote the latent effects of the J diplotypes. The model we propose is

$$y_i | \boldsymbol{\beta}, \mathbf{b}, \mathbf{c}_i, \sigma_e^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_i \boldsymbol{\beta} + \mathbf{c}_i \mathbf{b}, \sigma_e^2), i = 1, \dots, n.$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Sigma})$$

$$\sigma_e^2 \sim IG(a_e, b_e) \qquad\qquad (4.2)$$

$$\boldsymbol{\pi} \sim Dirichlet(\boldsymbol{\alpha})$$

where IG denotes the inverse gamma distribution, μ_0, Σ_0, a_e, b_e are assumed to be known. To complete the model, we also need a prior on $\mathbf{b}$. We consider two different priors. The first option is to use a standard linear mixed model, where b_j has a single normal prior distribution.

$$b_j | \sigma_b^2 \overset{\text{i.i.d.}}{\sim} N(0, \sigma_b^2), j = 1, \dots, J$$

$$\sigma_b^2 \sim IG(a_1, b_1) \qquad\qquad (4.3)$$

The second option assumes random cluster effects b_j follow a discrete distribution G, where G has a Dirichlet process prior(DPP) [9, 6, 30]. Additional details on the semiparametic approach we describe are given in [23]. The model is written formally as:

$$[b_j] \sim G$$

$$[G | M, \Phi_0] \sim DP(M \times G_0(\Phi_0)) \qquad\qquad (4.4)$$

where G_0 is called the base measure and represents a distribution that approximates the true nonparametric shape of G. The positive scaler M reflects our prior belief about how similar G is to the base measure G_0. For computational purposes, we select $G_0 = N(0, \sigma_b^2)$.

This second option allows for the possibility that each of the J clusters may not have distinct effects on the response. In that case, this second option allows the

posterior for $\mathbf{b}$ to take on fewer than J distinct values. We use simulation to explore the practical effects of these two priors for $\mathbf{b}$ in Section 4.3.

Clusters in our setting are defined as sets of individuals who carry a common pair of haplotypes, called diplotypes, across one or more genes. To understand how clusters can be formulated, consider the simple setting where there are two biallelic SNPs with alleles A_1/A_2 and B_1/B_2 respectively. In this case there are four haplotypes within a population, given by $h_1 = (A_1, B_1)$, $h_2 = (A_1, B_2)$, $h_3 = (A_2, B_1)$ and $h_4 = (A_2, B_2)$ at a given gene. Each cluster is comprised of individuals with one of the ten possible diplotypes, given by $D_1 = (h_1, h_2)$, $D_2 = (h_1, h_3)$, $D_3 = (h_1, h_4)$, $D_4 = (h_2, h_3)$, $D_5 = (h_2, h_4)$, $D_6 = (h_3, h_4)$, $D_7 = (h_1, h_1)$, $D_8 = (h_2, h_2)$, $D_9 = (h_3, h_3)$ and $D_{10} = (h_4, h_4)$. Alternative groupings of individuals is also tenable and can incorporate prior knowledge about the underlying genetic model. For example, if h_4 is known to be a dominant haplotype, then we can additionally group individuals with diplotypes D_3, D_5, D_6 and D_{10} together. Clusters can also be defined based on the presence of a single haplotype. In this case, each individual would belong to two clusters according to the corresponding pair of haplotypes. Now consider an individual who is heterozygous at both sites, so that the observed genotype is $(A_1 A_2, B_1 B_2)$. This individual is ambiguous between the two possible haplotype pairs (h_1, h_4) and (h_2, h_3). In fact, we will have haplotype ambiguity for all individuals who are heterozygous at at least two SNPs within a gene. If the number of SNPs is large, then there are many ways that this can occur. Since clusters are defined based on haplotype information, a similar level of ambiguity exists among cluster identifiers. In another words, the cluster membership is potentially unobserved for some individuals.

4.2.2 Estimation

Since it is difficult to sample the parameters from this joint posterior density, we use a well-described application of the Gibbs sampling as an approximation. Gibbs sampling [15, 13]is a form of Markov chain Monte Carlo simulation [33, 19]. It has been found very helpful in many multidimensional problems. Further details of this approach can be found in [14] and [16]. Briefly, the idea behind the Gibbs sampling is

that we are able to generate data from a joint posterior distribution of interest based on repeated sampling from a series of conditional distributions.

Based on our model in Equation 4.2 and Equation 4.3, the joint posterior density of all of the parameters under single normal prior is defined as:

$$Pr(\beta, b, \sigma_b^2, \sigma_e^2, \pi | y, \mathbf{c}) \propto IG(\sigma_b^2 | a_1, b_1) IG(\sigma_e^2 | a_e, b_e) N(\beta | \mu_0, \Sigma_0) \prod_{j=1}^{J} \pi_j^{\alpha_j - 1}$$

$$\times \prod_{i=1}^{n} N(y_i | x_i \beta + c_i^T b, \sigma_e^2) \prod_{j=1}^{J} N(b_j | 0, \sigma_b^2) \prod_{j=1}^{J} \pi_j^{\sum_{i=1}^{n} c_{ij}}$$

$$(4.5)$$

It is straightforward to verify that the conditional distributions for β, b_j, σ_b^2, σ_e^2 and π are given by:

$$[\beta | b, \sigma_b^2, \sigma_e^2, \pi, y, \mathbf{c}] \sim N(Ta, T)$$

$$[b_j | \beta, b_{-j}, \sigma_b^2, \sigma_e^2, \pi, y, \mathbf{c}] \sim N\left(\frac{\sigma_b^2}{\sigma_e^2 + \sum_{i=1}^{n} c_{ij} \sigma_b^2} \sum_{i=1}^{n} (y_i - x_i \beta) c_{ij}, \frac{\sigma_e^2 \sigma_b^2}{\sigma_e^2 + \sum_{i=1}^{n} c_{ij} \sigma_b^2} \right)$$

$$[\sigma_b^2 | \beta, b, \sigma_e^2, \pi, y, \mathbf{c}] \sim IG\left(a_1 + \frac{1}{2} J, b_1 + \frac{1}{2} \sum_{j=1}^{J} b_j^2 \right)$$

$$[\sigma_e^2 | \beta, b, \sigma_b^2, \pi, y, \mathbf{c}] \sim IG\left(a_e + \frac{1}{2} n, b_e + \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i \beta - c_i^T b)^2 \right)$$

$$[\pi | \beta, b, \sigma_b^2, \sigma_e^2, y, \mathbf{c}] \sim D(\alpha + \sum_{i=1}^{n} c_i)$$

$$(4.6)$$

where $T = \left(\sum_{i=1}^{n} x_i^T x_i / \sigma_e^2 + \Sigma_0^{-1} \right)^{-1}$, $a = \sum_{i=1}^{n} x_i^T (y_i - c_i^T b) / \sigma_e^2 + \Sigma_0^{-1} \mu_0$ and b_{-j} is the vector given by $(b_1, \ldots, b_{j-1}, b_{j+1}, \ldots, b_J)$.

Under the assumption of Dirichlet process prior for $\mathbf{b}$, conditional distributions for $\mathbf{b}$ is changed to:

$$[b_j | \beta, b_{-j}, \sigma_b^2, \sigma_e^2, \pi, y,] \propto q_k \delta_{b_k} + M q_0 \times N(0, \sigma_b^2) \prod_{i=1}^{n} [N(x_i \beta + b_j, \sigma_e^2)]^{I_{c_{ij}=1}} \quad (4.7)$$

where $q_k = \prod_{i=1}^{n} [N(x_i \beta + b_k, \sigma_e^2)]^{I_{c_{ij}=1}}$ and $q_0 = \int \prod_{i=1}^{n} [N(x_i \beta + b, \sigma_e^2)]^{I_{c_{ij}=1}} N(0, \sigma_b^2) db$.

Notice the above conditional distributions involve c_{ij}. Since the cluster identifiers are potentially unobservable, c_{ij} is missing for some i. Therefore, we propose multiply imputing the c_{ij} according to the posterior probability of group membership at the beginning of each iteration of the sampler. This posterior distribution is defined for each i and is conditional on the observed data and the most recent sample of the parameters from the posterior. Formally, for individual i, the posterior probability of membership to the jth group is given by:

$$Pr(c_{ij} = 1 | y, \beta, b, \sigma_e^2, \pi) = \frac{Pr(y_i | c_{ij} = 1) Pr(c_{ij} = 1) I[j \in \mathcal{S}(G_i)]}{\sum_{j=1}^{J} Pr(y_i | c_{ij} = 1) Pr(c_{ij} = 1) I[j \in \mathcal{S}(G_i)]} \quad (4.8)$$

where $Pr(y_i | c_{ij} = 1) = N(x_i \beta + b_j, \sigma_e^2)$, $Pr(c_{ij} = 1) = \pi_j$ and $I[j \in \mathcal{S}(G_i)] = 1$ if group j is in $\mathcal{S}(G_i)$ and 0 otherwise. Note that when c_i is observed, imputation of c_i is not necessary. This is summarized as follows:

The Gibbs sampler for drawing from the posterior distribution, $Pr(\beta, b, \sigma_b^2, \sigma_e^2, \pi | y, \mathbf{c})$, begins by selecting starting values $\beta^{(0)}$, $b^{(0)}$, $\sigma_b^{2(0)}$, $\sigma_e^{2(0)}$ and $\pi^{(0)}$ and setting $t = 0$. The sampler proceeds as follows:

ALGORITHM 1

1. For all i and j, impute c_{ij} according to Equation 4.8

2. Sample $\beta^{(t+1)}$ from $[\beta|b^{(t)}, \sigma_b^{2(t)}, \sigma_e^{2(t)}, \pi^{(t)}, y, \mathbf{c}]$.

3. Sample $\sigma_e^{2(t+1)}$ from $[\sigma_e^2|\beta^{(t+1)}, b^{(t)}, \sigma_b^{2(t)}, \pi^{(t)}, y, \mathbf{c}]$.

4. Sample $\sigma_b^{2(t+1)}$ from $[\sigma_b^2|\beta^{(t+1)}, b^{(t)}, \sigma_e^{2(t+1)}, \pi^{(t)}, y, \mathbf{c}]$.

5. Sample $\pi^{(t+1)}$ from $[\pi|\beta^{(t+1)}, b^{(t)}, \sigma_b^{2(t+1)}, \sigma_e^{2(t+1)}, y, \mathbf{c}]$.

6. Sample $b = (b_1, \ldots, b_J)$ as follows:

   - Sample $b_1^{(t+1)}$ from $[b_1|\beta^{(t+1)}, b_{-1}^{(t)}, \sigma_b^{2(t+1)}, \sigma_e^{2(t+1)}, \pi^{(t+1)}, y, \mathbf{c}]$.

     $\vdots$

   - Sample $b_J^{(t+1)}$ from $[b_J|\beta^{(t+1)}, b_{-J}^{(t+1)}, \sigma_b^{2(t+1)}, \sigma_e^{2(t+1)}, \pi^{(t+1)}, y, \mathbf{c}]$.

7. Set $t = t + 1$ and repeat steps (1)-(6) until convergence is met.

4.2.3 Convergence assesment

In practice, we use the multiple-chain diagnostic method described in [14] to check for convergence of the algorithm. Briefly, this method involves first generating m independent Gibbs sampling sequences of length n. Parameters are initialized with over dispersed values and the Gibbs sampler is run for each set of initial values. Then the simulations from the second halves of all the sequences together are collected. For each scaler estimand ψ, let ψ_{ij} be the ith simulation in the jth sequence. This method monitor convergence by estimating the factor by which the scale of the current distribution for ψ might be reduced if the simulations were continued in the limit $n \longrightarrow \infty$. This potential scale reduction $\hat{R}$ is estimated by:

$$\hat{R} = \sqrt{\frac{\hat{var}^+(\psi|y)}{W}}$$

where

$$\hat{var}^+(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

and B and W are the between- and within-sequence variances. They are computed by the following equation.

56

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2, \text{where} \bar{\psi}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}, \bar{\psi}_{\cdot\cdot} = \sum_{j=1}^{m}\bar{\psi}_{\cdot j}$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j{}^2, \text{where} s_j{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}n(\psi_{ij} - \bar{\psi}_{\cdot j})^2 \qquad (4.9)$$

If $\hat{R}$ is not near 1 for all of the parameters, continue the simulation. If $\hat{R}$ near 1 for all scaler estimands of interest, treat the selected draws as samples from the target distribution and take summary statistics for our estimates.

4.3 Simulation Study

In order to evaluate our methods described in Section 4.2, we conduct a simulation study. We start by first assuming the random effects are coming from a single normal prior, and estimation using 2 approaches described in this manuscript. A detailed simulation result is presented in Table 4.1 for varying ratios of σ_b and σ_e. In this simulation, a sample size of n = 1000 is assumed. The data is simulated assuming 5% ambiguity, 50 of the $n = 1000$ observations are ambiguous between the first two clusters. In all cases, 21 clusters are assumed and frequencies ranging from 0.01 to 0.08. The simulation starts by assuming 21 clusters, each cluster has an assumed cluster frequency, with the sum of these frequencies equals to 1. Then for each individual, randomly assign the cluster membership according to these frequencies. The random effects are generated from a normal distribution with normal 0 and standard deviation ranging from 0.2 to 0.8. The average standard error(se), cover rate, and the average confidence interval length are reported. Coverage rate is defined as the percentage of simulations for which the true parameter value is within the 95% confidence interval. Convergence is evaluated using the method described in Section 4.2.3. The results shows that both methods work well when σ_b/σ_e is 0.4 or higher. However, when $\sigma_b/\sigma_e = 0.2$, the coverage rate for cluster effect b is relatively low, but the CI length and standard error is smaller under dpp model compared with

a single normal prior model. Secondly, we assume the random effects are from a discrete model, and estimate the parameters assuming single normal prior and DPP model. Specifically, we assume 10 clusters in the data, and the first 4 cluster effects are some positive number, in our case, 2, 4, 8. The last 6 cluster effects are assumed to be 0. The simulation result is presented in Table 4.2. Surprisingly, both models works well. As expected, the bias and CI-length for ambiguous random effects b_{ua} are slightly higher under single normal prior model than DPP model.

Figure 4.1 illustrates the mean of the estimated cluster effect for cluster 1 and cluster 5 under DPP model.

Figure 4.1. Estimated Random Effects
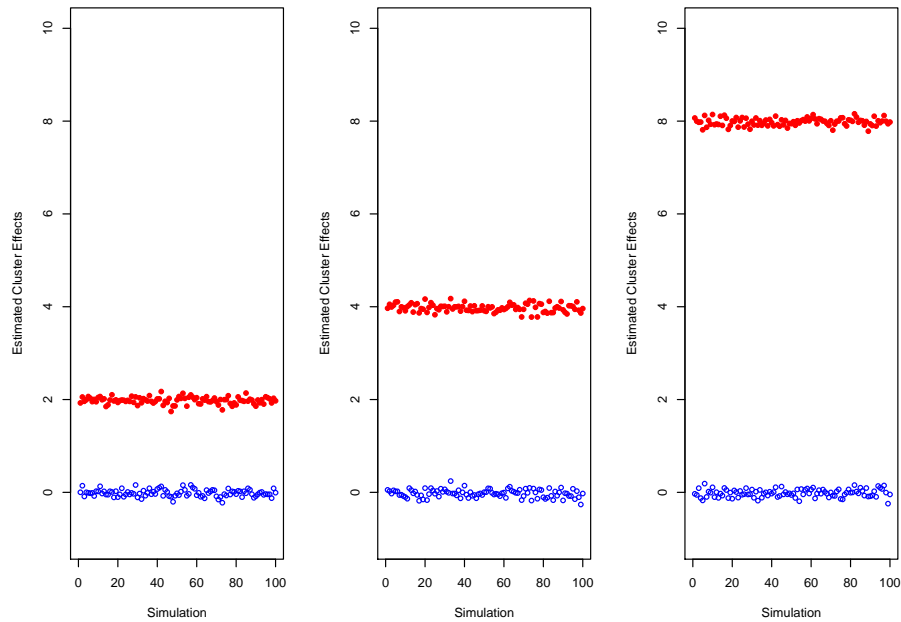
4.4 Discussion

In this manuscript, we describe a Bayesian approach to fitting the linear mixed model in the context of unobservable cluster identifiers. Two priors for cluster effects are proposed. We first assume the cluster effect follows a single normal distribution; then we relax this assumption and assume the cluster effect follows a Dirichlet process

Table 4.1. Simulation Results for differing variance ratios

σ_b/σ_e	Coverage rates[†]							Bias[‡]						
	$\bar{\beta}$	$\bar{b}_a$	$\bar{b}_{ua}$	$\bar{\pi}_a$	$\bar{\pi}_{ua}$	σ_b	σ_e	$\bar{\beta}$	$\bar{b}_a$	$\bar{b}_{ua}$	$\bar{\pi}_a$	$\bar{\pi}_{ua}$	σ_b	σ_e
0.2	0.97	0.97	0.96	0.93	0.96	0.94	0.96	0.02	0.11	0.11	0.01	0	0.06	0.02
0.4	0.96	1	1	0.92	0.96	0.94	0.96	0.04	0.15	0.14	0.01	0	0.06	0.02
0.6	0.96	1	1	0.96	0.96	0.94	0.91	0.05	0.16	0.16	0.01	0	0.08	0.02
0.8	0.96	1	1	0.93	0.96	0.96	0.95	0.07	0.18	0.19	0.01	0	0.12	0.02

σ_b/σ_e	CI-Length(se)[*]						
	$\bar{\beta}$	$\bar{b}_a$	$\bar{b}_{ua}$	$\bar{\pi}_a$	$\bar{\pi}_{ua}$	σ_b	σ_e
0.2	.12 (0.03)	0.62 (0.16)	0.62 (0.16)	0.04 (0.01)	0.03 (0.01)	0.32 (0.08)	0.09 (0.02)
0.4	0.22 (0.06)	1.45 (0.37)	1.43 (0.36)	0.04 (0.01)	0.03 (0.01)	0.31 (0.08)	0.09 (0.02)
0.6	0.28 (0.07)	2.1 (0.54)	2.19 (0.56)	0.04 (0.01)	0.03 (0.01)	0.41 (0.1)	0.1 (0.03)
0.8	0.38 (0.09)	3.02 (0.77)	3.02 (0.77)	0.04 (0.01)	0.03 (0.01)	0.56 (0.14)	0.09 (0.02)

(a) The cluster effects are generated assuming a single normal prior. Estimation using single normal prior model.

σ_b/σ_e	Coverage rates[†]							Bias[‡]						
	$\bar{\beta}$	$\bar{b}_a$	$\bar{b}_{ua}$	$\bar{\pi}_a$	$\bar{\pi}_{ua}$	σ_b	σ_e	$\bar{\beta}$	$\bar{b}_a$	$\bar{b}_{ua}$	$\bar{\pi}_a$	$\bar{\pi}_{ua}$	σ_b	σ_e
0.2	0.97	0.63	0.62	0.91	0.96	–	0.89	0.02	0.15	0.15	0.01	0	–	0.02
0.4	0.96	1	0.99	0.91	0.95	–	0.92	0.04	0.15	0.15	0.01	0	–	0.02
0.6	0.97	1	1	0.94	0.96	–	0.94	0.06	0.16	0.17	0.01	0	–	0.02
0.8	0.96	1	1	0.98	0.96	–	0.96	0.06	0.18	0.18	0	0	–	0.02

σ_b/σ_e	CI-Length(se)[*]						
	$\bar{\beta}$	$\bar{b}_a$	$\bar{b}_{ua}$	$\bar{\pi}_a$	$\bar{\pi}_{ua}$	σ_b	σ_e
0.2	0.13 (0.03)	0.31 (0.08)	0.32 (0.08)	0.04 (0.01)	0.03 (0.01)	–	0.09 (0.02)
0.4	0.19 (0.05)	1.25 (0.32)	1.31 (0.33)	0.04 (0.01)	0.03 (0.01)	–	0.09 (0.02)
0.6	0.29 (0.07)	2.12 (0.54)	2.16 (0.55)	0.03 (0.01)	0.03 (0.01)	–	0.09 (0.02)
0.8	0.34 (0.09)	2.87 (0.73)	2.94 (0.75)	0.04 (0.01)	0.03 (0.01)	–	0.08 (0.02)

(b) The cluster effects are generated assuming a single normal prior. Estimation using Drichlet process prior (DPP) model.

[†]Coverage rate is defined as the proportion of simulations for which the true parameter value is within the corresponding 95% confidence interval; [‡]Bias is defined as the absolute difference between the mean of the estimate over the simulations and the true parameter value; [*]CI-Length(se)is defined as the lenth of the 95% confidence interval and the standard error of the parameter estimates; $\bar{\beta}$ denotes average over all βs; $\bar{b}_a$ and $\bar{b}_{ua}$ denote average over all random effects bs across ambiguous and unambiguous clusters. $\bar{\pi}_a$ and $\bar{\pi}_{ua}$ denote average over all cluster frequencies πs across ambiguous and unambiguous clusters.

Table 4.2. Simulation Results for differing random effects

σ_b/σ_e	COVERAGE RATES[†]						BIAS[‡]					
	$\bar\beta$	$\bar b_a$	$\bar b_{ua}$	$\bar\pi_a$	$\bar\pi_{ua}$	σ_e	$\bar\beta$	$\bar b_a$	$\bar b_{ua}$	$\bar\pi_a$	$\bar\pi_{ua}$	σ_e
2	0.9	0.93	0.93	0.95	0.95	0.95	0.01	0.23	0.25	0.02	0.01	0.04
4	0.9	0.94	0.95	0.94	0.95	0.93	0.01	0.21	0.24	0.02	0.01	0.04
8	0.93	0.94	0.94	0.95	0.96	0.97	0.01	0.19	0.25	0.02	0.01	0.04

σ_b/σ_e	CI-LENGTH(SE)[*]					
	$\bar\beta$	$\bar b_a$	$\bar b_{ua}$	$\bar\pi_a$	$\bar\pi_{ua}$	σ_e
2	0.06 (0.01)	0.9 (0.23)	1.15 (0.29)	0.11 (0.03)	0.07 (0.02)	0.18 (0.05)
4	0.06 (0.01)	0.96 (0.24)	1.21 (0.31)	0.11 (0.03)	0.07 (0.02)	0.21 (0.05)
8	0.06 (0.01)	0.91 (0.23)	1.25 (0.32)	0.1 (0.03)	0.07 (0.02)	0.18 (0.05)

(a) The cluster effects are generated from a discrete distribution, with the b_1 through b_4 equals a positive number from 2 to 8. Estimation using single normal prior model.

σ_b/σ_e	COVERAGE RATES[†]						BIAS[‡]					
	$\bar\beta$	$\bar b_a$	$\bar b_{ua}$	$\bar\pi_a$	$\bar\pi_{ua}$	σ_e	$\bar\beta$	$\bar b_a$	$\bar b_{ua}$	$\bar\pi_a$	$\bar\pi_{ua}$	σ_e
2	0.9	0.91	0.93	0.96	0.95	0.97	0.01	0.22	0.2	0.02	0.01	0.05
4	0.94	0.93	0.94	0.97	0.94	0.95	0.01	0.19	0.18	0.02	0.01	0.04
8	0.93	0.95	0.95	0.95	0.94	0.95	0.01	0.2	0.17	0.02	0.01	0.04

σ_b/σ_e	CI-LENGTH(SE)[*]					
	$\bar\beta$	$\bar b_a$	$\bar b_{ua}$	$\bar\pi_a$	$\bar\pi_{ua}$	σ_e
2	0.06 (0.01)	0.98 (0.25)	0.95 (0.24)	0.1 (0.03)	0.07 (0.02)	0.22 (0.06)
4	0.06 (0.02)	0.93 (0.24)	0.92 (0.23)	0.1 (0.03)	0.07 (0.02)	0.21 (0.05)
8	0.06 (0.01)	0.98 (0.25)	0.9 (0.23)	0.11 (0.03)	0.07 (0.02)	0.19 (0.05)

(b) The cluster effects are generated from a discrete distribution, with the b_1 through b_4 equals a positive number from 2 to 8. Estimation using DPP model. See Figure 1 legend for definitions of terms.

prior(DPP). Gibbs sampler is used to arrive at the estimates. To address the ambiguity of the cluster identifiers, we impute the cluster membership according to the posterior probability of belonging to each cluster at beginning of each iteration.

CHAPTER 5

CONCLUSION

Three methods have been described in my thesis.

In chapter 2, I described a novel model fitting approach to arriving at maximum likelihood estimates of haplotype frequencies in a population of children multiply infected with the parasite that causes malaria. This approach offers two primary advantages over existing methods. First, the computational efficiency of our algorithm allows us to characterize a large number of sites. Secondly, our method also allows for a variable number of clones within an individual, making it more flexible than approaches designed for diploid populations.

In chapter 3, I described a method for estimation and test the haplotype effect on the disease phenotype. Inferring the haplotype effect is important because the association between the disease phenotype and the haplotypes is likely to provide more information on the complex relationship between genetic variation and phenotype than any single SNP can provide. Characterizing the association between polymorphisms in the parasite genome and measured traits in an infected human host may provide insight into disease etiology while ultimately informing new strategies for improved treatment and prevention. The method I proposed provides a comprehensive framework for simultaneous estimation of population haplotype frequencies and haplotype-trait associations in a general setting where the number of clones within an individual is variable.

In chapter 4, I described two Bayesian approaches for estimation and testing in the context of mixed modeling with missing cluster identifiers. In the method we proposed, two prior distributions are assumed for cluster effects. First we assume a single normal prior. Then we relax this assumption and instead assume a Dirichlet

process prior. Gibbs sampler are used for iteratively arriving at estimation. In order to account for the unknown cluster identifier, we propose to impute the cluster membership for each individual at the beginning of each iteration.

APPENDIX A

FIXED ASSUMPTION ON THE NUMBER OF INFECTIONS

Note that the sum of the population level haplotype frequencies must equal 1, so we have $\theta_K = 1 - \sum_{k=1}^{K-1} \theta_k$. Equation 3.9 is then given by:

$$\frac{\partial E\left[\log L(\Phi|X^{(com)})|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right]}{\partial \theta_k} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \left(\frac{\delta_{ik}}{\theta_k} - \frac{\delta_{iK}}{1 - \sum_{k=1}^{K-1} \theta_k}\right) = 0$$

for $k = 1, \ldots, K-1$, or equivalently,

$$\begin{bmatrix} \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{i1}}{\theta_1} \\ \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{i2}}{\theta_2} \\ \vdots \\ \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{iK-1}}{\theta_{K-1}} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{iK}}{1 - \sum_{k=1}^{K-1} \theta_k} \\ \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{iK}}{1 - \sum_{k=1}^{K-1} \theta_k} \\ \vdots \\ \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{iK}}{1 - \sum_{k=1}^{K-1} \theta_k} \end{bmatrix} \tag{A.1}$$

Note that all of the elements of the vector in the right hand of the above equation are equal. Therefore, we can set the first element of the vector in the left hand of Equation A.1 equal to each of the remaining elements of this vector. That is, for $k = 2, \ldots, K-1$, we have:

$$\frac{\sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{i1}}{\theta_1} = \frac{\sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})\delta_{ik}}{\theta_k}$$

or equivalently:

$$\theta_k = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik}}{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1}} \theta_1 \tag{A.2}$$

Thus we can derive an estimate of θ_1 and then use Equation A.2 to find estimates of θ_k for $k = 2, \ldots, K-1$. To find $\widehat{\theta}_1$, note we can write:

$$1 - \sum_{k=1}^{K-1} \theta_k = 1 - \theta_1 - \theta_1 \sum_{k=2}^{K-1} \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik}}{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1}}$$

Therefore from the first element of Equation A.1, we have:

$$\theta_1 \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{iK}$$

$$= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1} \left(1 - \theta_1 - \theta_1 \sum_{k=2}^{K-1} \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik}}{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1}} \right)$$

$$= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1} - \theta_1 \sum_{k=1}^{K-1} \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik}$$

Equivalently:

$$\theta_1 \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \sum_{k=1}^{K} \delta_{ik} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1} \tag{A.3}$$

Note $\sum_{k=1}^{K} \delta_{ik} = C$ and $\sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i} = 1$ and so Equation A.3 yields:

$$\widehat{\theta}_1^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1}}{nC}$$

65

APPENDIX B

POISSON ASSUMPTION ON THE NUMBER OF INFECTIONS

Under the Poisson assumption, we have $\sum_{k=1}^{K} \delta_{ik} = c_i$ and therefore Equation A.3 is written:

$$\theta_1 \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) c_i = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1}$$

resulting in

$$\widehat{\theta}_1^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{i1}}{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) c_i}$$

APPENDIX C

SEMI-PARAMETRIC ASSUMPTION

Note that the sum of the q_c must equal 1, so we have $q_C = 1 - \sum_{c=1}^{C-1} q_c$ and Equation 3.16 is given by:

$$\frac{\partial E\left[\log L(\Phi|X^{(com))}|\mathbf{X}^{(obs)}, \widehat{\Phi}^{(t)}\right]}{\partial q_c} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \left[\frac{I(c_i = c)}{q_c} - \frac{I(c_i = C)}{1 - \sum_{c=1}^{C} q_c}\right] = 0$$

for $c = 1, \ldots, C - 1$. This is equivalent to:

$$
\begin{bmatrix}
\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i=1)}{q_1} \\
\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i=2)}{q_2} \\
\vdots \\
\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i=C-1)}{q_{C-1}}
\end{bmatrix}
=
\begin{bmatrix}
\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i=C)}{1 - \sum_{c=1}^{C} q_c} \\
\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i=C)}{1 - \sum_{c=1}^{C} q_c} \\
\vdots \\
\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i=C)}{1 - \sum_{c=1}^{C} q_c}
\end{bmatrix}
\tag{C.1}
$$

Since all of the elements of the vector on the right hand side of the above equation are equal, we can set each element of the vector on the left hand side equal to first element of this vector. That is, we can write:

$$\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = 1)}{q_1} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = c)}{q_c}$$

for $c = 2, \ldots, C - 1$. Equivalently, we have:

$$q_c = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = c)}{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = 1)} q_1 \tag{C.2}$$

Thus, similar to the estimation of θ in Appendix A.1, we can derive an estimate of q_1 and then use Equation C.2 to find estimates of q_c for $c = 2, \ldots, C - 1$. To find $\widehat{q}_1$, note we can write:

$$1 - \sum_{c=1}^{C-1} q_c = 1 - q_1 - q_1 \sum_{c=2}^{C-1} \frac{\sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = c)}{\sum_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = 1)}$$

and using the same approach as we did for deriving Equation A.3, we have:

$$q_1 \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \sum_{c=1}^{C} I(c_i = c) = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = 1) \tag{C.3}$$

Since $\sum_{c=1}^{C} I(c_i = c) = 1$ and $\sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i} = 1$ Equation C.3 yields:

$$\widehat{q}_1^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) I(c_i = 1)$$

APPENDIX D

ESTIMATION IN QUASI-SPECIES SETTING

From Equation 3.24, we have:

$$\frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik}}{\alpha_k} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)})(1 - \delta_{ik})}{(1 - \alpha_k)}$$

or equivalently:

$$\alpha_k \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik} \tag{D.1}$$

Since $\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) = n$, Equation D.1 yields:

$$\widehat{\alpha}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\widehat{\Phi}^{(t)}) \delta_{ik}}{n}$$

BIBLIOGRAPHY

[1] Benjamini, Yoav, and Hochberg, Yosef. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological 57* (1995), 289–300.

[2] Benjamini, Yoav, and Yekutieli, Daniel. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*, 4 (2001), 1165–1188.

[3] Breman, J. G. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg. 64(1-2) S* (2001), 1–11.

[4] Clark, A. G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol 7*, 2 (1990), 111–122.

[5] Dempster, A. P., Larid, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39(1)* (1977), 1–38.

[6] Escobar, M.D. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association 89*, 425 (1994), 268–277.

[7] Escobar, M.D., and West, M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 430 (1995), 577–588.

[8] Excoffier, L., and Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol. 12* (1995), 921–927.

[9] Ferguson, R.S. A bayesian analysis of some non-parammetric problems. *The annals of Statistics 1* (1973), 209–230.

[10] Foulkes, and DeGruttola, V. A resampling-based approach to multiple testing with uncertainty in phase. *International Journal of Biostatistics 3* (2007).

[11] Foulkes, A.S., Yucel, R., and Li, X. A likelihood based approach to mixed modeling with ambiguity in cluster identifiers. *Biostatistics in press* (2008).

[12] Foulkes, A.S., Yucel, R., and Reilly, M. Mixed modeling and multiple imputation for unobservable genotype clusters. *Statistics in Medicine 10.1002/sim.3051* (2007).

[13] Gelfand, A.E., and Smith, A.F.M. Sampling-based approaches to calculating marginal densities. *J. American Statistical Association 85*, Suppl1 (1990), 398–409.

[14] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. *Bayesian Data Analysis.* Chapman and Hall, 2004.

[15] Gelman, S., and Gelman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *EEE Transactions on Pattern Analysis and Machine Intelligence 6*, Suppl1 (1984), 721–741.

[16] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice.* Chapman and Hall, 1996.

[17] Greenwood, B. M., Bojang, K., Whitty, C. J., and Targett, G. A. Malaria. *Lancet 365* (2005), 1487–1498.

[18] Guerra, C.A., Gikandi, P.W., Tatem, A.J., Noor, A.M., Smith, D.L., Hay, S.I., and Snow, R.W. The limits and intensity of plasmodium falciparum transmission: Implications for malaria control and elimination worldwide. *PLoS Medicine 5*, 2 (2008), e38.

[19] Hastings, WK. Monte carlo sampling-based methods using markov chains and their applications. *Biometrika 57*, Suppl1 (1970), 97–109.

[20] Hill, W. G., and Babiker, H. A. Estimation of number of malaria clones in blood samples. *Proceedings of the Royal Society of London 262* (1995), 249–257.

[21] Hollingdale, M.R., Nardin, E.H., Tharavanij, S., Schwartz, A.L., and Nussenzweig, R.S. Inhibition of entry of Plasmodium falciparum and P. vivax sporozoites into cultured cells; an in vitro assay of protective antibodies. *J Immunol 132*, 2 (1984), 909–913.

[22] Ito, T., Chiku, S., Inoue, E., Tomita, M., Morisaki, T., Morisaki, H., and Kamatani, N. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am. J. Hum. Genet. 72(2)* (2003), 384–398.

[23] Kleinman, K.P., and Ibrahim, J.G. A semiparametric bayesian approach to the random effects model. *Biometrics 54* (1998), 921–938.

[24] Laird, N. M., and Ware, J.H. Random-effects models for longitudinal data. *Biometrics 38* (1982), 963–974.

[25] Lake, S. L., Lyon, H., Tantisira, K., Silverman, E. K., Weiss, S. T., Laird, N. M., and Schaid, D. J. Estimation and testing of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity 55* (2003), 56–65.

[26] Li, X., Foulkes, A.S., Yucel, R., and Rich, S. An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. *Statistical Application in Genetics and Molecular Biology 6* (2007).

[27] Lin, D.Y., and Zeng, D. Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association 101*, 473 (2006), 89–104.

[28] Lin, S., Cutler, D. J., Zwick, M. E., and Chakravarti, A. Haplotype inference in random population samples. *Am. J. Hum. Genet. 71* (2002), 1129–1137.

[29] Louis, T. A. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B 44*, 2 (1982), 226–233.

[30] MacEachern, S.N. Estimating normal means with a conjugate style dirichlet process prior. *Comuunications in Statistics 23* (1994), 727–741.

[31] McLachlan, G. J., and Krishnan, T. *The EM Algorithm and Extensions.* John Wiley & Sons, 1997.

[32] Meilijson, Isaac. A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society 39* (1989), 1–38.

[33] Metropolis, N., W., Rosenbluthm A., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equations of state calculations by fast computing machines. *J. Chemical Physics 21*, Suppl1 (1953), 1087–1091.

[34] Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. Bayesian haplotype inference for multiple linked single -nucleotide polymorphisms. *Am J Hum Genet 70*, 1 (2002), 157–169.

[35] Pe'er, I., and Beckmann, J. S. Recovering frequencies of known haplotype blocks from single-nucleotide polymorphism allele frequencies. *Genetics 166* (2004), 2001–2006.

[36] Pollard, K.S., and van der Laan, M.J. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference 125* (2004), 85–100.

[37] Schaid, D., Rowland, C., Tines, D., Jacobson, R., and Poland, G. Score tests for association between traits and haplotypes when linkage phase in ambiguous. *Am. J. Hum. Genet. 70* (2002), 425–34.

[38] Schumacher, F.R., and Kraft, P. A bayesian latent class analysis for whole-genome association analyses: an illustration using the gaw15 simulated rheumatoid arthritis dense scan data. *BMC Proceedings 1*, Suppl1 (2007), S112.

[39] Stephens, M., and Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet. 73* (2003), 1162–9.

[40] Stephens, M., Smith, N. J., and Donnelly, P. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics 68* (2001), 978–989.

[41] Storey, J.D., and Tibshirani, R. Statistical significance for genomewide studies. *PNAS 100*, 16 (2003), 9440–9445.

[42] Wang, S., Kidd, K. K., and Zhao, H. On the use of DNA pooling to estimate haplotype frequencies. *Genet Epidemiol 24* (2003), 74–82.

[43] Westfall, P.H., and Young, S.S. *Resampling-based multiple testing.* John Wiley & Sons, 1993.

[44] Zavala, F., Cochrane, A. H.and Nardin, E. H., Nussenzweig, R. S., and Nussenzweig, v. Circumsporozoite proteins of malaria parasites contain a single immunodominant region with two or more identical epitopes. *J. Exp. Med. 157 (6)* (1983), 1947–1957.

[45] Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J., and Ehm, M.G. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity 53* (2002), 79–91.