Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2009)
58, *Part* 5, *pp.* 663–678

Estimating and testing haplotype–trait associations in non-diploid populations

X. Li,

University of Massachusetts, Amherst, USA

B. N. Thomas,

Rochester Institute of Technology, USA

S. M. Rich and D. Ecker,

University of Massachusetts, Amherst, USA

J. K. Tumwine

Makerere University, Kampala, Uganda

and A. S. Foulkes

University of Massachusetts, Amherst, USA

[Received June 2008. Final revision January 2009]

Summary. Malaria is an infectious disease that is 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 aetiology and ultimately inform 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, which is 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 human immunodeficiency virus quasi-species setting. A simulation study is presented to characterize performance of the method. 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.

Keywords: Expectation–maximization algorithm; Genotype; Haplotype; Human immunodeficiency virus; Linear model; Malaria; Phenotype; Quasi-species; Strain

1. Introduction

Our investigation is motivated by a study of the human pathogenic species *Plasmodium falci-*

Address for correspondence: A. S. Foulkes, School of Public Health and Health Sciences, University of Massachusetts, 404 Arnold House, 715 North Pleasant Street, Amherst, MA 01003, USA.
E-mail: foulkes@schoolph.umass.edu

 0035–9254/09/58663

parum, the group of parasites that cause malaria. Here, interest lies in characterizing associations between genetic polymorphisms in the haploid parasite and clinical measures of severity of disease, such as red blood cell (RBC) 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 multiple strains within a single human host. In general, the observed genotype data consist of the set of nucleotides 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 nucleotides on a single chromosome, which is called the haplotype, 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 is unknown and this number can vary across individuals. Combined, these two challenges serve as the motivation for our present research. Consider, for example, an individual who is infected by multiple parasites. Further suppose that the observed genotype for this individual is Aa at one site and Bb at a second site on the genome, where A, a, B and b are used to represent the observed nucleotide. In this simple case, there are four possible haplotypes: $h_1 = (A, B)$, $h_2 = (A, b)$, $h_3 = (a, B)$ and $h_4 = (a, b)$. The precise combination of these haplotypes within this individual is not observable. In a human population, the number of homologous chromosomes 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 haplotype 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 two, three or four 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.

Several methods for characterizing population level haplotype frequencies and haplotype–trait associations in human populations have been described (Excoffier and Slatkin, 1995; Stephens *et al.*, 2001; Zaykin *et al.*, 2002; Stephens and Donnelly, 2003; Schaid *et al.*, 2002; Lake *et al.*, 2003; Lin and Zeng, 2006; Foulkes *et al.*, 2008). In this paper, we propose an extension of the EM approach for haplotype–trait association studies (Lake *et al.*, 2003; Lin and Zeng, 2006) 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.

In previous work, we described an expectation–maximization (EM) type of algorithm for estimating haplotype frequencies in the malaria setting that uses only the observed genotype data (Li *et al.*, 2007). This prior work, while extending the methods of Excoffier and Slatkin (1995) and Hill and Babiker (1995), does not take into account phenotypic or clinical information about the host. In this paper, 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. Thus the method that is presented represents an extension of Li *et al.* (2007) to incorporate trait information as well as an extension of Lake *et al.* (2003) and Lin and Zeng (2006) 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 that is presented herein provides the framework for evaluating these potential associations.

In Section 2, 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, 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 4 describes extensions for the human immunodeficiency virus (HIV) quasi-species setting in which multiple strains can arise from repeat infections though, more generally, this is a result of external pressures, such as treatment exposures. Finally, in Section 5 we provide a discussion of our findings.

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 involve different assumptions about the distribution of the number of infections:

(a) we assume that the number of infections within a host is fixed at a constant $C > 0$;
(b) we assume that this number follows a conditional Poisson distribution where we condition on the presence of at least one infection;
(c) 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.

2.1. Notation

Let $\mathbf{G} = (G_1, \ldots, G_n)$ where G_i is the unphased (observed) multisite genotype for individual i. Further suppose that $\boldsymbol{\mathcal{H}} = (\mathcal{H}_1, \ldots, \mathcal{H}_n)$ 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 $\mathbf{Y}$ by using the generalized linear model such that the expected value of Y_i is related to the linear predictor $(\mathbf{X}_i^{\mathrm{T}} \quad \mathbf{H}_i^{\mathrm{T}})\boldsymbol{\beta}$ through a link function g:

$$g(E[Y_i]) = (\mathbf{X}_i^{\mathrm{T}} \quad \mathbf{H}_i^{\mathrm{T}})\boldsymbol{\beta} \tag{1}$$

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, $g(\cdot)$ reduces to the identity link. Since the haplotype combination for individual i is potentially unobserved, we consider all possible $\mathcal{H}_i$ that are consistent with the observed genotype data, as described in Section 2.2. $\mathbf{H}_i$ 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 Lin and Zeng (2006).

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 generalized linear model framework, we assume that the probability density of $\mathbf{Y}$ is from an exponential family, given by

$$\Pr(\mathbf{Y}|\mathbf{X},\mathbf{H},\boldsymbol{\beta}) = L(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\mathbf{H}) = \prod_{i=1}^{n} \exp\left[\frac{Y_i(\mathbf{X}_i^{\mathrm{T}} \quad \mathbf{H}_i^{\mathrm{T}})\boldsymbol{\beta} - b\{(\mathbf{X}_i^{\mathrm{T}} \quad \mathbf{H}_i^{\mathrm{T}})\boldsymbol{\beta}\}}{a(\psi)} + c(Y_i,\psi)\right] \quad (2)$$

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 a missing data problem and thus an EM-type algorithm is a natural choice for this setting. The EM algorithm, which was formalized by Dempster *et al.* (1977), 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 has been met. In our setting, the observed data consist of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{G}$ and are denoted $\mathbf{X}^{(\mathrm{obs})}$, whereas the complete data consist of $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{G}$ and $\mathcal{H}$ and are denoted $\mathbf{X}^{(\mathrm{com})}$. 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

$$L(\Phi|\mathbf{X}^{(\mathrm{com})}) = \prod_{i=1}^{n} \Pr(Y_i|\mathbf{X}_i,\mathbf{H}_i,\boldsymbol{\beta}) \Pr(\mathcal{H}_i|\boldsymbol{\theta}) \quad (3)$$

where $\Pr(\mathcal{H}_i|\boldsymbol{\theta})$ is the corresponding haplotype set probabilities for the ith individual. Notably, this likelihood assumes that the haplotype frequencies are independent of environmental or 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 that are made with respect to the number of infections.

Let $\hat{\Phi}^{(t)}$ 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

$$E[\log\{L(\Phi|\mathbf{X}^{(\mathrm{com})})\}|\mathbf{X}^{(\mathrm{obs})},\hat{\Phi}^{(t)}] = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})[\log\{\Pr(Y_i|\mathbf{X}_i,\mathbf{H}_i,\boldsymbol{\beta})\} + \log\{\Pr(\mathcal{H}_i|\boldsymbol{\theta})\}] \quad (4)$$

where

$$p_{iH_i}(\hat{\Phi}^{(t)}) = p\{\mathcal{H}_i|\mathcal{H}_i \in \mathcal{S}(G_i), Y_i, \mathbf{X}_i, \hat{\Phi}^{(t)}\} = \frac{\Pr(Y_i|\mathbf{X}_i,\mathbf{H}_i,\hat{\boldsymbol{\beta}}^{(t)}) \Pr(\mathcal{H}_i|\hat{\boldsymbol{\theta}}^{(t)})}{\sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} \Pr(Y_i|X_i,H_i,\hat{\boldsymbol{\beta}}^{(t)}) \Pr(\mathcal{H}_i|\hat{\boldsymbol{\theta}}^{(t)})}. \quad (5)$$

Next, we maximize the conditional expectation of the complete-data log-likelihood given in equation (4). It is straightforward to show that the $(t+1)$th estimate of Φ can be obtained by finding the root for the equations

$$\begin{aligned}\frac{\partial E[\log\{L(\Phi|\mathbf{X}^{(\mathrm{com})})\}|\mathbf{X}^{(\mathrm{obs})},\hat{\Phi}^{(t)}]}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \frac{\partial \log\{L(\boldsymbol{\beta}|Y_i,\mathbf{X}_i,\mathbf{H}_i)\}}{\partial\boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \frac{(Y_i - E[Y_i|\mathbf{X}_i,\mathbf{H}_i,\boldsymbol{\beta}])\,(\mathbf{X}_i^{\mathrm{T}} \quad \mathbf{H}_i^{\mathrm{T}})^{\mathrm{T}}}{a(\psi)} \\ &= 0 \end{aligned} \quad (6)$$

and

$$\frac{\partial E[\log\{L(\Phi|\mathbf{X}^{(\text{com})})\}|\mathbf{X}^{(\text{obs})}, \hat{\Phi}^{(t)}]}{\partial \theta_k} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \frac{\partial \log\{\Pr(\mathcal{H}_i|\boldsymbol{\theta})\}}{\partial \theta_k} = 0. \tag{7}$$

As noted in Lake *et al.* (2003) for the diploid setting, equation (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 convergence of the algorithm when $\max(|\hat{\Phi}^{(t)} - \hat{\Phi}^{(t+1)}|/\hat{\Phi}^{(t)}) < 1.0 \times 10^{-5}$. Alternatively, a convergence criterion can be based on the observed data likelihood, which is given by

$$\prod_{i=1}^{n} \sum_{H_i \in S(G_i)} \Pr(Y_i|\mathbf{X}_i, \mathbf{H}_i, \boldsymbol{\beta}) \Pr(H_i|\theta).$$

2.2.1. Fixed number of infections

Let δ_{ik} denote the number of copies of haplotype h_k in the haplotype combination $\mathcal{H}_i$. First suppose that there are exactly C strains in each individual where $C > 0$, i.e. assume that each individual has exactly C infections, where C is some known positive integer. This implies that $\Sigma_{k=1}^{K} \delta_{ik} = C$, where δ_{ik} ranges from 1 to C. $\Pr(\mathcal{H}_i|\boldsymbol{\theta})$ of equation (3) is thus given by

$$\Pr(\mathcal{H}_i|\boldsymbol{\theta}) = \frac{C!}{\delta_{i1}! \ldots \delta_{iK}!} \prod_{k=1}^{K} \theta_k^{\delta_{ik}}. \tag{8}$$

In this case, $\Phi = (\boldsymbol{\beta}, \boldsymbol{\theta})$. Plugging equation (8) into equation (7), we have

$$\frac{\partial E[\log\{L(\Phi|\mathbf{X}^{(\text{com})})\}|\mathbf{X}^{(\text{obs})}, \hat{\Phi}^{(t)}]}{\partial \theta_k} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \left(\frac{\delta_{ik}}{\theta_k} - \frac{\delta_{iK}}{\theta_K} \right) = 0. \tag{9}$$

Resulting closed form solutions for $\hat{\theta}_k$ (see Appendix A.1) are given by

$$\hat{\theta}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \delta_{ik}}{nC}. \tag{10}$$

2.2.2. Poisson assumption on the numbers of infections

In Section 2.2.1 we assumed 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 Hill and Babiker (1995). Since data sets are generally comprised only of individuals with at least one detectable infection, the conditional Poisson model is considered. Let the Poisson model conditioning on at least one infection be given by

$$\phi_c(\lambda) = \begin{cases} \dfrac{\lambda^c/c!}{\exp(\lambda) - 1} & c > 0, \\ 0 & c = 0 \end{cases} \tag{11}$$

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 (8) for the haplotype combination probabilities is now replaced by

$$\Pr(\mathcal{H}_i|\boldsymbol{\theta},\lambda)=\Pr(\mathcal{H}_i,c_i|\boldsymbol{\theta},\lambda)=\phi_{c_i}(\lambda)\frac{c_i!}{\delta_{i1}!\ldots\delta_{iK}!}\prod_{k=1}^{K}\theta_k^{\delta_{ik}} \tag{12}$$

where c_i is the number of infections for the ith individual. Estimation of $\boldsymbol{\theta}$ proceeds similarly to the setting in which C is fixed. Straightforward calculation (see Appendix A.2) leads to closed form solutions for $\hat{\theta}_k$ given by

$$\hat{\theta}_k^{(t+1)}=\frac{\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)}p_{iH_i}(\hat{\Phi}^{(t)})\delta_{ik}}{\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)}p_{iH_i}(\hat{\Phi}^{(t)})c_i}. \tag{13}$$

Estimation of λ is achieved by solving

$$\begin{aligned}\frac{\partial E[\log\{L(\Phi|\mathbf{X}^{(\mathrm{com})})\}|\mathbf{X}^{(\mathrm{obs})},\hat{\Phi}^{(t)}]}{\partial\lambda}&=\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)}p_{iH_i}(\hat{\Phi}^{(t)})\,\frac{\partial\log\{\Pr(\mathcal{H}_i|\boldsymbol{\theta},\lambda)\}}{\partial\lambda}\\&=\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)}p_{iH_i}(\hat{\Phi}^{(t)})\left\{\frac{c_i}{\lambda}-\frac{\exp(\lambda)}{\exp(\lambda)-1}\right\}=0.\end{aligned} \tag{14}$$

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.

2.2.3. *Semiparametric 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 $\mathbf{q}=(q_1,\ldots,q_C)$ and $\Phi=(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{q})$. Equation (8) for the haplotype set probabilities is now replaced by

$$\Pr(\mathcal{H}_i|\boldsymbol{\theta},\mathbf{q})=\frac{c_i!}{\delta_{i1}!\ldots\delta_{iK}!}\prod_{k=1}^{K}\theta_k^{\delta_{ik}}\prod_{c=1}^{C}q_c^{I(c_i=c)} \tag{15}$$

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

$$\begin{aligned}\frac{\partial E[\log\{L(\Phi|\mathbf{X}^{(\mathrm{com})})\}|\mathbf{X}^{(\mathrm{obs})},\hat{\Phi}^{(t)}]}{\partial q_c}&=\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)}p_{iH_i}(\hat{\Phi}^{(t)})\frac{\partial\log\{\Pr(H_i|\boldsymbol{\theta},\mathbf{q})\}}{\partial q_c}\\&=\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)}p_{iH_i}(\hat{\Phi}^{(t)})\left\{\frac{I(c_i=c)}{q_c}-\frac{I(c_i=C)}{q_C}\right\}=0\end{aligned} \tag{16}$$

and resulting closed form solutions (see Appendix A.3) for $\hat{q}_c$ are given by

$$\hat{q}_c^{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)}p_{iH_i}(\hat{\Phi}^{(t)})\,I(c_i=c). \tag{17}$$

2.3. *Inference*

Wald tests are used to test hypotheses of haplotype–trait associations. 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's method within the EM framework (Louis, 1982). An alternative approach is to approximate the observed information matrix with the empirical observed information matrix which can be computed by (Meilijson, 1989)

$$I_e(\Phi; \mathbf{X}) = \sum_{i=1}^{n} s_i(\Phi) s_i^{\mathrm{T}}(\Phi)|_{\Phi=\hat{\Phi}} \tag{18}$$

where $\hat{\Phi}$ is the estimate of the parameters in the last EM iteration and $s_i(\Phi)$ is the score function from the observed data likelihood for the ith individual. The score is given by (McLachlan and Krishnan, 1997)

$$s_i(\Phi) = E_\Phi \left[\frac{\partial \log\{L_i(\Phi|X_i^{(\mathrm{com})})\}}{\partial \Phi} \middle| X_i^{(\mathrm{obs})}, \hat{\Phi} \right]. \tag{19}$$

For example, under the fixed number of infections assumption, we have

$$s_i(\Phi) = \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}) \begin{pmatrix} (Y_i - E[Y_i|\mathbf{X}_i, \mathbf{H}_i, \beta]) \, (\mathbf{X}_i^{\mathrm{T}} \quad \mathbf{H}_i^{\mathrm{T}})^{\mathrm{T}} / a(\psi) \\ \delta_{i1}/\theta_1 - \delta_{iK}/\theta_K \\ \vdots \\ \delta_{ik-1}/\theta_{k-1} - \delta_{iK}/\theta_K \end{pmatrix}. \tag{20}$$

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 (1) is set equal to the identity link and we have the linear regression model

$$Y_i = (\mathbf{X}_i^{\mathrm{T}} \quad \mathbf{H}_i^{\mathrm{T}}) \, \boldsymbol{\beta} + \varepsilon_i. \tag{21}$$

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 $\mathbf{X}$ and $\mathbf{H}$. In the real data example that is 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.

3.1. Simulation study

To evaluate the performance of the methods that were described in Section 2, we conduct a simulation study and report the type 1 error rates (ERs) and power under each of the three models for the number of infections: fixed, Poisson and semiparametic. The simulation starts by generating the number of infections c for each individual. Under the fixed number 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 semiparametric 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 on the basis of the multinomial distribution. Four haplotypes, which are 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 by 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, $B = 200$ data sets with sample sizes of $n = 500$ are generated. Analysis is performed using genotype data and trait information only, i.e. we assume that the haplotypic phase and the number of infections are unknown and apply the methods that were described in Section 2.

Simulation results are provided in Table 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 $\widehat{\text{se}}$. 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 less than 0.001 to 0.086 and the coverage rates are between 0.92 and 0.97. This suggests that our algorithm 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 greater than 80% power to detect moderate effect sizes of greater than 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 number 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.

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 by using the fixed approach, assuming that 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 2, part (a). Comparing this with correct application of the semiparametric method (Table 1), we see a dramatic loss of power 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 $\beta_1 \geqslant 0.4$. Secondly, we perform estimation by using the fixed number approach, again assuming that 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 2, part (b). Comparing these results with correct application of the Poisson approach with $\lambda = 2$ (Table 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 $\beta \geqslant 0.2$. These findings support the use of the more sophisticated modelling approaches in these settings.

Next, we perform estimation by 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 we present the results in Table 2, part (c). Here the modelling approach provides estimates of λ and, from this, we calculate $\hat{q}_c$ as $(\hat{\lambda}^c/c!)/\{\exp(\hat{\lambda}) - 1\}$. 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 1). Finally, we evaluate performance in applying the semiparametric approach when the number of infections actually arises from a Poisson distribution with $\lambda = 2$.

Table 1. Simulation results for the dominant model under three assumptions†

	β_1‡	*Bias* ($\widehat{se}$)§			*Coverage rates*§§			*Power**	*ER***
		$\hat{\beta}_1$	$\hat{\lambda}$	$\bar{\bar{\theta}}$	β_1	λ	$\bar{\theta}$		
Fixed number model									
$C=2$	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
	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
$C=3$	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
	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
$C=4$	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
	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 model									
$\lambda=2$	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
	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
$\lambda=3$	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
	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
			$\bar{\bar{q}}$			$\bar{q}$			
Semi-parametric model									
	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

†$\bar{\bar{\theta}}$ and $\bar{\bar{\mathbf{q}}}$ denote averaging across all $\hat{\theta}$s and $\hat{q}$s respectively. $\bar{\theta}$ and $\bar{\mathbf{q}}$ denote averaging across all θs and qs respectively.
‡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 1 error rate.

These results are given in Table 2, part (d), and, as expected, we see a slight loss of 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 1). Power for the semiparametric approach is estimated to be 0.81. Since we are not incorporating knowledge about the distribution of the number of infections the loss of power is expected.

Table 2. Sensitivity analysis to model misspecification

β_1^*	*Bias*			*Coverage rates*			*Power*	*ER*
	$\hat{\beta_1}$	$\bar{\hat{\mathbf{q}}}$	$\bar{\hat{\boldsymbol{\theta}}}$	β_1	$\bar{\mathbf{q}}$	$\bar{\boldsymbol{\theta}}$		
(a) Incorrect application of the fixed approach under semiparametric data†								
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
(b) Incorrect application of the fixed approach under Poisson-distributed data‡								
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
(c) Incorrect application of the conditional Poisson model§								
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
(d) Incorrect application of the semiparametric approach under Poisson-distributed data§§								
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

†The data are simulated assuming between one and five infections with equal probabilities of 0.20 whereas the estimation approach assumes $c = 2$ fixed infections. See the caption for Fig. 1 for definitions of terms.
‡The data are simulated assuming a conditional Poisson distribution with $\lambda = 2$, whereas the estimation procedure assumes $c = 2$ fixed infections.
§The data are simulated assuming between one and five infections with equal probabilities of 0.20.
§§The data are simulated assuming a conditional Poisson distribution with $\lambda = 2$. The number of infections is assumed to range from 1 to 10.

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 (Breman, 2001). The majority of these deaths are in children under the age of 5 years and in resource-constrained settings since current treatment options are costly or unavailable (Greenwood *et al.*, 2005; Guerra *et al.*, 2008). Recent advances in sequencing technologies provide new opportunities for population-based genetic association studies to uncover complex relationships between genetic polymorphisms and measures of progression of disease. 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 an EM approach (see Section 2) to data arising from a cross-sectional study of $n = 126$ malaria-infected children from Uganda. We focus on haplotypes in one polymorphic circumsporozoite protein (CSP) region (CSP-TH3) of the gene that encodes for a cellular adhesion domain of the CSP. The CSP facilitates adhesion of the parasite to liver cells, which is a critical initial step in its replication process in a human host (Zavala *et al.*, 1983; Hollingdale *et al.*, 1984). The goal of our analysis is to uncover haplotype associations with RBC count (log-transformed). The RBC count is a well-known diagnostic tool for detecting anaemia, which is a common and often lethal manifestation of malaria.

Data on 12 sites, 10 of which are polymorphic in our sample, are considered. Notably, sites that are constant across our data do not inform the analysis but are included for completeness. Across all individuals, we see up to three different nucleotides at a site and, within a single individual, one or two 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 Li *et al.* (2007). 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 by using the approach of Li *et al.* (2007). We assume a Poisson distribution and apply the approach of Section 2.2.2. A dominant genetic model is assumed, as in the simulation study.

Estimated haplotype effects on the RBC and corresponding p-values for tests of the null hypotheses that these effects equal 0 are provided in Table 3. The p-values are unadjusted for multiple comparisons. Using a Bonferroni adjustment, p-values that are less than $0.05/14 = 0.0036$ are considered significant at the 0.05-level. A significant association is observed between the RBC count and the three haplotypes numbered 8, 11 and 12. Interestingly, the effect of carrying at least one copy of haplotype 11 appears to increase the RBC count $\exp(0.344) = 1.41$-fold, suggesting a potential protective effect. In contrast, haplotypes 8 and 12 result in a lower RBC count with estimated decreases of $\exp(-0.484) = 0.616$-fold and $\exp(-0.137) = 0.872$-fold respectively.

Table 3. Estimated haplotype effects for Uganda†

	Unique haplotype	*Estimated frequency* $\hat{\theta}$	*Estimated effect* $(\hat{\beta})$	*Standard error*	*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

†The results are based on a sample of size $n = 126$ and assume a Poisson model for the number of strains per individual.
‡The haplotype effect on the RBC count is significantly different from 0 after applying a Bonferroni adjustment for multiple comparisons.

Notably, the estimated number of individuals with each of these haplotypes (which is given by $126\hat{\theta}_k$) is small and further confirmatory research is required to make firm conclusions.

4. Further extensions for the quasi-species setting

In the methods that were 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 who are infected with 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 a 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. Although this distinction is subtle, it does require modification of the estimation approach that was described in Section 2. Again let G_i be the unphased (observed) multisite genotype for the ith individual where $i = 1, \ldots, n$. Further suppose that $\mathcal{H}_i$ 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, i.e. 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

$$\delta_{ik} = \begin{cases} 1 & \text{if } h_k \text{ is present in the } i\text{th individual,} \\ 0 & \text{if } h_k \text{ is not present in the } i\text{th individual.} \end{cases} \tag{22}$$

Under the model that is given in equation (1), the complete likelihood function can again be written as in equation (3) where $\Pr(\mathcal{H}_i|\boldsymbol{\theta})$ is replaced with

$$\Pr(\mathcal{H}_i|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \alpha_k^{\delta_{ik}} (1-\alpha_k)^{1-\delta_{ik}}. \tag{23}$$

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 equation

$$\begin{aligned} \frac{\partial E[\log\{L(\Phi)|\mathbf{X}^{(\mathrm{com})}\}|\mathbf{X}^{(\mathrm{obs})}, \hat{\Phi}^{(t)}]}{\partial \alpha_k} &= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \frac{\partial \log\{\Pr(\mathcal{H}_i|\boldsymbol{\alpha})\}}{\partial \alpha_k} \\ &= \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \left(\frac{\delta_{ik}}{\alpha_k} - \frac{1-\delta_{ik}}{1-\alpha_k} \right) = 0. \end{aligned} \tag{24}$$

Resulting closed form solutions (see Appendix A.4) for $\hat{\alpha}_k$ are given by

$$\hat{\alpha}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \delta_{ik}}{n}. \tag{25}$$

5. Discussion

In this paper, we describe an approach to estimate and test haplotype–trait associations between individuals with multiple strains of an infectious agent. Three approaches to modelling the number of infections were described in Section 2. 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 Lake *et al.* (2003). 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 that is made in equation (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 ERs but substantial bias in these probability estimates. The semiparametric approach performs reasonably well under the Poisson model with a slight loss of power. Incorrect application of the fixed number approach leads to more substantial losses of power, reductions in coverage rates and increases in type 1 ERs. Applications of the correct models lead to reasonable power and control of type 1 ERs.

Coupled with this investigation is the need for appropriate methods for controlling type 1 ERs in the context of multiple comparisons. In Section 3.2, we applied a Bonferroni correction to assess significance. Alternative single-step and step-down methods that are based on the false discovery rate and that account for the correlated nature of these tests (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Storey and Tibshirani, 2003) are also tenable. In addition, further consideration of resampling-based approaches and related extensions (Westfall and Young, 1993; Pollard and van der Laan, 2004; Foulkes and DeGruttola, 2007) may be appropriate. Extensions of the mixed effects modelling approaches that were developed originally for the diploid setting (Foulkes *et al.*, 2007, 2008) 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 that were sufficiently severe to warrant at least one visit to the doctor. This is a potential limitation of the method that is 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 three 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 aetiology 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 that are 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.

Acknowledgements

Support for this research was provided by the National Institute of Allergy and Infectious Diseases research award R01AI056983. Funding for data collection was provided by the National Institute of General Medicine (R01GM070077) and the World Health Organization (TDR-A10375).

Appendix A

A.1. Estimation under fixed number of infections

Note that the sum of the population level haplotype frequencies must equal 1, so we have $\theta_K = 1 - \Sigma_{k=1}^{K-1}\theta_k$. Equation (9) is then given by

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

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

$$\begin{pmatrix} \sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{i1}/\theta_1 \\ \sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{i2}/\theta_2 \\ \vdots \\ \sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{i_{K-1}}/\theta_{K-1} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{iK} \Big/ \left(1-\sum_{k=1}^{K-1}\theta_k\right) \\ \sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{iK} \Big/ \left(1-\sum_{k=1}^{K-1}\theta_k\right) \\ \vdots \\ \sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{iK} \Big/ \left(1-\sum_{k=1}^{K-1}\theta_k\right) \end{pmatrix}. \tag{26}$$

Note that all the elements of the vector on the right-hand side of equation (26) are equal. Therefore, we can set the first element of the vector on the left-hand side of equation (26) equal to each of the remaining elements of this vector, which yields

$$\theta_k = \frac{\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{ik}}{\sum_{i=1}^{n}\sum_{\mathcal{H}_i\in\mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)})\delta_{i1}}\theta_1. \tag{27}$$

Thus we can derive an estimate of θ_1 and use equation (27) to find estimates of θ_k for $k = 2, \ldots, K-1$. From the first element of equation (26) we have

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

or equivalently

$$\theta_1 \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \sum_{k=1}^{K} \delta_{ik} = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \delta_{i1}. \tag{28}$$

Finally, since $\Sigma_{k=1}^{K} \delta_{ik} = C$ and $\Sigma_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i} = 1$, equation (28) yields

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

A.2. Estimation under Poisson assumption

Under the Poisson assumption, we have $\Sigma_{k=1}^{K} \delta_{ik} = c_i$ and therefore equation (28) is written

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

resulting in

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

A.3. Estimation for semiparametric approach

Note that the sum of the q_c must equal 1, so we have $q_C = 1 - \Sigma_{c=1}^{C-1} q_c$ and equation (16) is given by

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

for $c = 1, \ldots, C-1$. Using the same approach as described in Appendix A.1, we have

$$\hat{q}_c^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \, I(c_i = c). \tag{29}$$

A.4. Estimation for quasi-species setting

From equation (24), we have

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

or equivalently

$$\alpha_k \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) = \sum_{i=1}^{n} \sum_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) \delta_{ik}. \tag{30}$$

Since $\Sigma_{i=1}^{n} \Sigma_{\mathcal{H}_i \in \mathcal{S}(G_i)} p_{iH_i}(\hat{\Phi}^{(t)}) = n$, equation (30) yields

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

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

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

Foulkes, A. and DeGruttola, V. (2007) A resampling-based approach to multiple testing with uncertainty in phase. *Int. J. Biostatist.*, **3**, article 2.

Foulkes, A., Yucel, R. and Li, X. (2008) A likelihood based approach to mixed modeling with ambiguity in cluster identifiers. *Biostatistics*, **9**, 635–657.

Foulkes, A., Yucel, R. and Reilly, M. (2007) Mixed modeling and multiple imputation for unobservable genotype clusters. *Statist. Med.*, **27**, 2784–2801.

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

Guerra, C., Gikandi, P., Tatem, A., Noor, A., Smith, D., Hay, S. and Snow, R. (2008) The limits and intensity of plasmodium falciparum transmission: implications for malaria control and elimination worldwide. *PLoS Med.*, **5**, no. 2, e38.

Hill, W. G. and Babiker, H. A. (1995) Estimation of number of malaria clones in blood samples. *Proc. R. Soc. Lond.*, **262**, 249–257.

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

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

Li, X., Foulkes, A., Yucel, R. and Rich, S. (2007) An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. *Statist. Applic. Genet. Molec. Biol.*, **6**, article 33.

Lin, D. and Zeng, D. (2006) Likelihood-based inference on haplotype effects in genetic association studies. *J. Am. Statist. Ass.*, **101**, 89–104.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: Wiley.

Meilijson, I. (1989) A fast improvement to the EM algorithm on its own terms. *J. R. Statist. Soc.* B, **51**, 127–138.

Pollard, K. and van der Laan, M. (2004) Choice of a null distribution in resampling-based multiple testing. *J. Statist. Planng Inf.*, **125**, 85–100.

Schaid, D., Rowland, C., Tines, D., Jacobson, R. and Poland, G. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **70**, 425–434.

Stephens, M. and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Stephens, M., Smith, N. J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Storey, J. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natn. Acad. Sci. USA*, **100**, 9440–9445.

Westfall, P. and Young, S. (1993) *Resampling-based Multiple Testing*. New York: Wiley.

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

Zaykin, D., Westfall, P., Young, S., Karnoub, M., Wagner, M. and Ehm, M. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.*, **53**, 79–91.