

Haplotype Frequency Estimation in the Presence of Genotyping Errors

Guohua Zou¹, Hongyu Zhao^{1,2}

Departments of ¹Epidemiology & Public Health and ²Genetics,
Yale University School of Medicine, New Haven, CT

Corresponding author:

Hongyu Zhao, Ph.D.

Department of Epidemiology and Public Health

60 College Street

Yale University School of Medicine

New Haven, CT 06520-8034

Phone: (203) 785-6271

Fax: (203) 785-6912

Email: hongyu.zhao@yale.edu

Key Words: genotyping errors, haplotype frequency, population data, family data

Running Title: Haplotype frequency estimation

Abstract

Several statistical methods have been proposed to estimate haplotype frequencies, either based on unrelated individuals or based on families. These estimates may yield insights on population genetics as well as associations between candidate regions and disease of interest. One limitation of the existing methods is that all these methods make the implicit assumption that there are no genotyping errors. However, genotyping errors are unavoidable in practice. Numerous methods have been developed to incorporate genotyping errors in genetic studies, but none to date have addressed the issues of haplotype inference in the presence of genotyping errors. In this article, we develop statistical methods for haplotype inference incorporating genotyping errors. We describe how our methods can be applied to analyze unrelated individuals as well as nuclear families. Our simulation results show that the proposed methods perform well in the presence of genotyping errors.

Introduction

Genetic analyses using haplotype data may reveal more information than those based on single markers both in the study of population history and in the identification of genes associated with complex diseases (see, for example, Fallin et al. 2001). One issue that is often encountered in using haplotypes is that most genotyping technologies only lead to individual marker information, not haplotypes. Many statistical methods have been developed to estimate haplotype frequencies and reconstruct haplotype pairs in individuals, including Clark (1990), Excoffier and Slatkin (1995), Hawley and Kidd (1995), Long et al. (1995), Stephens et al. (2001) and Niu et al. (2002) for unrelated individuals and Excoffier and Slatkin (1998) and Schaid (2002) for nuclear families.

One implicit assumption in these methods is that there are no genotyping errors, although such errors are unavoidable and may affect many aspects of genetic data analysis, such as type-I error, power, and parameter estimation (Terwilliger et al. 1990; Buetow 1991; Shields et al. 1991; Goldstein et al. 1997; Gordon et al. 1999; Akey et al. 2001). Numerous statistical methods have been developed to incorporate genotyping errors in the analysis of genotype data (Broman and Weber 1998; Göring and Terwilliger 2000a, 2000b, 2000c, 2000d; Gordon and Ott 2001; Gordon et al. 2001; Sobel et al. 2002). However, no statistical method has been developed to incorporate genotyping errors in haplotype inference. In this article, we describe statistical methods for haplotype frequency estimation in the presence of genotyping errors, both for samples consisting of unrelated individuals and for samples consisting of nuclear families. We derive closed-form expressions for the maximum likelihood estimates (MLEs) under simple situations, e.g. one biallelic marker for unrelated individuals and for nuclear families with one child. For more complex situations, haplotype frequencies can be estimated using the expectation-maximization (EM) algorithm (Dempster et al. 1977). These haplotype frequency estimates allow us to estimate the number of genotype configurations taking into account of genotyping errors. These reconstructed genotype configuration counts

can be used to perform statistical analysis methods to identify associations between candidate genes and complex diseases. Simulation results show that our methods perform well in the presence of genotyping errors.

Methods

In this section, we describe our methods for haplotype frequency estimation for samples consisting of unrelated individuals and samples consisting of nuclear families in the presence of genotyping errors.

Notation

Let k be the number of genetic markers and I_j be the number of alleles at marker j , where $j = 1, \dots, k$. The alleles at marker j are denoted by $1, \dots, I_j$. We use $2 \times k$ matrix

$$\begin{pmatrix} a_{11} & \cdots & a_{k1} \\ a_{12} & \cdots & a_{k2} \end{pmatrix}$$

to denote the genotype of an individual, where a_{j1} and a_{j2} are the two alleles at marker j with $1 \leq a_{j1}, a_{j2} \leq I_j$ ($j = 1, \dots, k$). Denote the total number of possible haplotypes by H , which is equal to $\prod_{j=1}^k I_j$. The frequencies of these H haplotypes are denoted by p_1, \dots, p_H , respectively. Let the sample size be n . Throughout the paper, we assume Hardy-Weinberg equilibrium and random mating.

The case of one biallelic marker for unrelated individuals

The simplest case for unrelated individuals is that of only one biallelic marker, i.e. $k = 1$ and $I_1 = 2$. We assume that genotyping errors are independently introduced into each marker of each individual. The genotyping error rates from true allele 1 to erroneous allele 2 and from true allele 2 to erroneous allele 1 are ε_1 and ε_2 , respectively. Note that other types of errors have been described in the literature (see Sobel et al. 2002 for a detailed discussion). Let the numbers of the observed

three genotypes $(1\ 1)'$, $(1\ 2)'$ and $(2\ 2)'$ (here $'$ means transpose) be n_1 , n_2 and n_3 , where $n_1 + n_2 + n_3 = n$.

Let O denote the observed genotype, and T denote the true genotype. Then we have

$$\begin{aligned} P(O = (1\ 1)') &= P(O = (1\ 1)' | T = (1\ 1)') \cdot P(T = (1\ 1)') \\ &\quad + P(O = (1\ 1)' | T = (1\ 2)') \cdot P(T = (1\ 2)') \\ &\quad + P(O = (1\ 1)' | T = (2\ 2)') \cdot P(T = (2\ 2)') \\ &= [(1 - \varepsilon_1)p_1 + \varepsilon_2 p_2]^2, \end{aligned}$$

where p_1 is the frequency of allele 1, and $p_2 = 1 - p_1$. Similarly,

$$\begin{aligned} P(O = (1\ 2)') &= 2[(1 - \varepsilon_1)p_1 + \varepsilon_2 p_2][\varepsilon_1 p_1 + (1 - \varepsilon_2)p_2], \\ P(O = (2\ 2)') &= [\varepsilon_1 p_1 + (1 - \varepsilon_2)p_2]^2. \end{aligned}$$

Based on these probabilities, we can derive the MLE of allele frequency p_1 as

$$\hat{p}_1 = \frac{(2n_1 + n_2)/(2n) - \varepsilon_2}{1 - \varepsilon_1 - \varepsilon_2}. \quad (1)$$

Based on this estimate, we can estimate the numbers of genotypes $(1\ 1)'$, $(1\ 2)'$ and $(2\ 2)'$ as $n\hat{p}_1^2$, $2n\hat{p}_1(1 - \hat{p}_1)$ and $n(1 - \hat{p}_1)^2$, respectively.

The case of multiple markers and multiple alleles for unrelated individuals

For the general case of multiple markers and multiple alleles at each marker, for simplicity, we assume that the error rate from one allele to another allele is the same at the same marker and denote this error rate by ε_l for marker l , where $l = 1, \dots, k$. Instead of deriving closed-form expressions for haplotype frequency estimates, we use the EM algorithm to obtain these estimates. Denote the observed genotype for the i th individual by g_i , $i = 1, \dots, n$. Let $d_{i(u)}$, $u = 1, \dots, U_i$, denote all possible diplotypes which are consistent with genotype g_i . By diplotype, we mean the haplotype pair, that is, the genotype whose phase information is known. In this situation, we distinguish the parental origin of the two haplotypes among the U_i

compatible haplotype pairs, although it is impossible to do so based on unrelated individuals. For example, diplotypes $\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$ and $\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}$ are considered to be different, where the top haplotype can be considered having paternal origin and the bottom haplotype having maternal origin.

If the true diplotype is d for the i th individual, the probability that the observed genotype is g_i is given by

$$\begin{aligned}
P(g_i | d) &= \sum_{u=1}^{U_i} P(d_{i(u)} | d) \\
&= \sum_{u=1}^{U_i} \prod_{l=1}^k P(d_{i(u)}^{(l)} | d^{(l)}) \\
&= \sum_{u=1}^{U_i} \prod_{l=1}^k [1 - (I_l - 1)\varepsilon_l]^{z_{iuld}} \varepsilon_l^{2 - z_{iuld}}, \tag{2}
\end{aligned}$$

where $d_{i(u)}^{(l)}$ and $d^{(l)}$ denote the genotypes of $d_{i(u)}$ and d at the l th marker, and z_{iuld} denotes the number of the same alleles on the same chromosome for $d_{i(u)}^{(l)}$ and $d^{(l)}$, where z_{iuld} is 0, 1, or 2.

Suppose that the (unknown) diplotype of the i th individual is d_i , then the log-likelihood function is

$$\log L = \sum_{i=1}^n \log P(g_i, d_i).$$

The MLEs of haplotype frequencies can be obtained through the following EM algorithm.

(i) **E-step:**

Let $p = (p_1, \dots, p_H)$ and denote the estimates of p at the j th step by $p^{(j)}$. Then we have

$$\begin{aligned}
Q(p | p^{(j)}) &\equiv \sum_{i=1}^n E_{d_i | (g_i, p^{(j)})} [\log P(g_i, d_i)] \\
&= \sum_{i=1}^n \sum_d [\log P(g_i, d)] \cdot P(d | g_i, p^{(j)}) \\
&= \sum_{i=1}^n \sum_d [\log P(g_i | d) + \log P(d)] \cdot P(d | g_i, p^{(j)}), \tag{3}
\end{aligned}$$

where \sum_d means summation over all possible diplotypes, $P(g_i | d)$ is given by formula (2), and

$$\begin{aligned}
P(d | g_i, p^{(j)}) &= \frac{P(g_i | d) \cdot P(d | p^{(j)})}{P(g_i | p^{(j)})} \\
&= \frac{P(g_i | d) \cdot P(d | p^{(j)})}{\sum_d P(g_i | d) \cdot P(d | p^{(j)})}.
\end{aligned}$$

(ii) **M-step:**

Note that when $d = \begin{pmatrix} h_s \\ h_t \end{pmatrix}$, where h_s and h_t are the two haplotypes for diplotype d , $P(d) = p_s^2$ if $s = t$ and $= 2p_s p_t$ otherwise under Hardy-Weinberg equilibrium. So

$$\frac{\partial \log P(d)}{\partial p_t} = \frac{\delta_{dt}}{p_t},$$

where δ_{dt} is equal to the number of times that haplotype h_t is present in diplotype d . δ_{dt} can be 0, 1, or 2, and $\sum_{t=1}^H \delta_{dt} = 2$. Because $\log P(g_i | d)$ does not depend on haplotype frequencies p_1, \dots, p_H , by maximizing the expectation function $Q(p | p^{(j)})$ given in (3), we can obtain the $(j + 1)$ th iteration value as follows:

$$\begin{aligned}
p_t^{(j+1)} &= \frac{\sum_{i=1}^n \sum_d \delta_{dt} \cdot P(d | g_i, p^{(j)})}{\sum_{t=1}^H \sum_{i=1}^n \sum_d \delta_{dt} \cdot P(d | g_i, p^{(j)})} \\
&= \frac{1}{2n} \sum_{i=1}^n \sum_d \delta_{dt} \cdot P(d | g_i, p^{(j)}). \tag{4}
\end{aligned}$$

Starting with some initial value $p^{(0)} = (p_1^{(0)}, \dots, p_H^{(0)})$ and using formula (4), we can iterate the E-step and M-step above until convergence to obtain MLEs of haplotype frequencies. Based on haplotype frequency estimates, we can obtain the estimates of the numbers of genotypes g_i^* :

$$n_i^* = n \cdot P(g_i^*) = n \cdot \sum_d^* P(d),$$

where \sum_d^* means summation over all different diplotypes consistent with genotype g_i^* .

The case of one biallelic marker for family trios

The simplest case for nuclear families is that of a biallelic marker for nuclear families with two parents and one offspring, i.e. $k = 1$, $I_1 = 2$, and the number of children $r = 1$. There are ten possible true trio genotypes (see the second column of Table 1), where the first two genotypes denote those of parents, the last genotype denotes that of child, and we make no distinction between parent 1 and parent 2. Let ε_1 and ε_2 be the genotyping error rates defined before in the discussion of single markers for unrelated individuals. Then there are 18 possible observed genotypes (see the fourth and last columns of Table 1). Eight of the 18 family genotypes are not Mendelian consistent. Although we can discard these trios showing inconsistent Mendelian inheritance, there is useful information on diplotypes in these genotypes that can be used to recover diplotypes of individuals if genotyping errors are appropriately taken into account. This can be achieved under the likelihood framework, and we can use all the observed data to estimate haplotype frequencies and the numbers of true trio genotypes T_i ($i = 1, \dots, 10$) in the sample as follows.

Suppose that the number of observed trio genotype O_i in the sample is n_i , $i = 1, \dots, 18$. Then the log-likelihood is given by (neglecting constants)

$$\log L = \sum_{i=1}^{18} n_i \log [P(O = O_i)], \quad (5)$$

where the probability $P(O = O_i)$ can be calculated by

$$P(O = O_i) = \sum_{m=1}^{10} P(O = O_i | T = T_m) \cdot P(T = T_m). \quad (6)$$

To illustrate the calculation of $P(O = O_i)$, we calculate $P(O = O_5)$ in the following. It can be seen that

$$\begin{aligned} & P(O = O_5 | T = T_8) \\ &= P(O = ((1 \ 1)' (1 \ 2)' (1 \ 2)') | T = ((1 \ 2)' (2 \ 2)' (1 \ 2)')) \\ &= 3(1 - \varepsilon_1)^2 \varepsilon_2^2 (1 - \varepsilon_2)^2 + 4 \varepsilon_1 (1 - \varepsilon_1) \varepsilon_2^3 (1 - \varepsilon_2) + \varepsilon_1^2 \varepsilon_2^4. \end{aligned}$$

Similarly, we can get the other conditional probabilities $P(O = O_5 | T = T_m)$. Substituting these formulas into (6) and using the values of $P(T = T_m)$ given in the third column of Table 1, we have

$$\begin{aligned} P(O = O_5) &= 8\varepsilon_1^2(1 - \varepsilon_1)^4 \cdot p_1^4 + \psi_1(\varepsilon_1, \varepsilon_2) \cdot 2p_1^3p_2 + \psi_2(\varepsilon_1, \varepsilon_2) \cdot 4p_1^2p_2^2 \\ &\quad + \psi_3(\varepsilon_1, \varepsilon_2) \cdot 2p_1p_2^3 + 8\varepsilon_2^4(1 - \varepsilon_2)^2 \cdot p_2^4, \end{aligned}$$

where

$$\begin{aligned} \psi_1(\varepsilon_1, \varepsilon_2) &= 2\varepsilon_1(1 - \varepsilon_1)^4(1 - \varepsilon_2) + 6\varepsilon_1^2(1 - \varepsilon_1)^3 \varepsilon_2 + (1 - \varepsilon_1)^4(1 - \varepsilon_2)^2 \\ &\quad + 4\varepsilon_1(1 - \varepsilon_1)^3 \varepsilon_2(1 - \varepsilon_2) + 3\varepsilon_1^2(1 - \varepsilon_1)^2 \varepsilon_2^2, \\ \psi_2(\varepsilon_1, \varepsilon_2) &= 2(1 - \varepsilon_1)^3 \varepsilon_2(1 - \varepsilon_2)^2 + 4\varepsilon_1(1 - \varepsilon_1)^2 \varepsilon_2^2(1 - \varepsilon_2) + 2\varepsilon_1^2(1 - \varepsilon_1) \varepsilon_2^3 \\ &\quad + (1 - \varepsilon_1)^2 \varepsilon_2^2(1 - \varepsilon_2)^2 + \varepsilon_1(1 - \varepsilon_1) \varepsilon_2^3(1 - \varepsilon_2), \end{aligned}$$

and

$$\begin{aligned} \psi_3(\varepsilon_1, \varepsilon_2) &= 6(1 - \varepsilon_1) \varepsilon_2^3(1 - \varepsilon_2)^2 + 2\varepsilon_1 \varepsilon_2^4(1 - \varepsilon_2) + 3(1 - \varepsilon_1)^2 \varepsilon_2^2(1 - \varepsilon_2)^2 \\ &\quad + 4\varepsilon_1(1 - \varepsilon_1) \varepsilon_2^3(1 - \varepsilon_2) + \varepsilon_1^2 \varepsilon_2^4. \end{aligned}$$

To simplify our calculation, we can use the following result described in Zou et al. (2003):

Lemma (i) Let $P(O = M_0 | T = M) = \phi(\varepsilon_1, \varepsilon_2)$. Then (a) $P(O = \underline{M}_0 | T = M) = \phi(1 - \varepsilon_1, 1 - \varepsilon_2)$; (b) $P(O = M_0 | T = \underline{M}) = \phi(1 - \varepsilon_2, 1 - \varepsilon_1)$.

(ii) Let $P(T = M_0 | O = M_0) = \varphi(p_1, \varepsilon_1, \varepsilon_2)$. Then $P(T = \underline{M}_0 | O = \underline{M}_0) = \varphi(1 - p_1, \varepsilon_2, \varepsilon_1)$,

where \underline{M} means the conjugate of genotype M . For a family with genotype M , we define the conjugate of M , denoted by \underline{M} , as the genotype with each 1 in M replaced by 2 and each 2 in M replaced by 1.

By maximizing the log-likelihood given by formula (5), we can find the MLE of p_1 . Using the allele frequency estimates, we can estimate the true trio genotype data: the estimated number of the true trio genotype T_i is the sample size multiplied by the estimated population frequency of T_i which can be obtained from the third column of Table 1 and the estimates of p_1 and p_2 . For example, the estimated number of the true trio genotype T_1 is $m_1 = n \cdot \hat{p}_1^4$.

The case of multiple markers and multiple alleles for nuclear families with multiple children

Similar to the discussion of multiple markers and multiple alleles for unrelated individual, we assume that the error rate is ε_l at the l th marker for general nuclear families. We also apply the EM algorithm to obtain the MLEs of haplotype frequencies as follows.

Let the observed genotype for the i th family be $(g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}})$, $i = 1, \dots, n$, where g_{if} , g_{im} and g_{ic_j} ($j = 1, \dots, n_i$) denote the genotypes of father, mother and children, respectively, and n_i denotes the number of children in the i th family. If the true diploypes of father, mother and children for i th family are d_f, d_m and $d_{c_1}, \dots, d_{c_{n_i}}$, respectively, then the probability that the observed genotype for the i th family is $(g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}})$ is given by

$$\begin{aligned}
\Phi_i &\equiv P(g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}} \mid d_f, d_m, d_{c_1}, \dots, d_{c_{n_i}}) \\
&= P(g_{if} \mid d_f) \cdot P(g_{im} \mid d_m) \cdot \prod_{v=1}^{n_i} P(g_{ic_v} \mid d_{c_v}) \\
&= \left[\sum_{u=1}^{U_{if}} P(d_{if(u)} \mid d_f) \right] \left[\sum_{u=1}^{U_{im}} P(d_{im(u)} \mid d_m) \right] \left\{ \prod_{v=1}^{n_i} \left[\sum_{u=1}^{U_{ic_v}} P(d_{ic_v(u)} \mid d_{c_v}) \right] \right\} \\
&= \left[\sum_{u=1}^{U_{if}} \prod_{l=1}^k P(d_{if(u)}^{(l)} \mid d_f^{(l)}) \right] \left[\sum_{u=1}^{U_{im}} \prod_{l=1}^k P(d_{im(u)}^{(l)} \mid d_m^{(l)}) \right] \\
&\quad \times \left\{ \prod_{v=1}^{n_i} \left[\sum_{u=1}^{U_{ic_v}} \prod_{l=1}^k P(d_{ic_v(u)}^{(l)} \mid d_{c_v}^{(l)}) \right] \right\},
\end{aligned}$$

where

$$\begin{aligned}
P(d_{if(u)}^{(l)} \mid d_f^{(l)}) &= [1 - (I_l - 1)\varepsilon_l]^{z_{ifuld_f}} \varepsilon_l^{2 - z_{ifuld_f}}, \\
P(d_{im(u)}^{(l)} \mid d_m^{(l)}) &= [1 - (I_l - 1)\varepsilon_l]^{z_{imuld_m}} \varepsilon_l^{2 - z_{imuld_m}}, \\
P(d_{ic_v(u)}^{(l)} \mid d_{c_v}^{(l)}) &= [1 - (I_l - 1)\varepsilon_l]^{z_{ic_vuld_{c_v}}} \varepsilon_l^{2 - z_{ic_vuld_{c_v}}},
\end{aligned}$$

and $d_{if(u)}$, $d_{if(u)}^{(l)}$, U_{if} , and z_{ifuld_f} , etc. are similarly defined as $d_{i(u)}$, $d_{i(u)}^{(l)}$, U_i , and z_{iuld} above. For example, $d_{ic_v(u)}$ denotes the u th diplotype of the v th child in the i th family which is consistent with genotype g_{ic_v} , and $z_{ic_vuld_{c_v}}$ denotes the number of the same alleles on the same chromosome for $d_{ic_v(u)}^{(l)}$ and $d_{c_v}^{(l)}$ which can take either 0, or 1 or 2 as its value. As in the case of unrelated individuals, we emphasize here that for (and only for) U_{if} ($U_{im}, U_{ic_v}, v = 1, \dots, n_i$) diploypes $d_{if(u)}$ ($d_{im(u)}$, $d_{ic_v(u)}, v = 1, \dots, n_i$) of genotype g_{if} ($g_{im}, g_{ic_v}, v = 1, \dots, n_i$), we distinguish different chromosomes.

Now we suppose that the (unknown) diplotype of the i th family is $(d_{if}, d_{im}, d_{ic_1}, \dots, d_{ic_{n_i}})$. Then the log-likelihood is

$$\log L = \sum_{i=1}^n \log P(g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}}, d_{if}, d_{im}, d_{ic_1}, \dots, d_{ic_{n_i}}).$$

The EM algorithm to obtain haplotype frequency estimates is performed as follows.

(i) **E-step:**

$$Q(p | p^{(j)}) \equiv \sum_{i=1}^n E_{(d_{if}, d_{im}, d_{ic_1}, \dots, d_{ic_{n_i}}) | (g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}}, p^{(j)})} [\log P(g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}}, d_{if}, d_{im}, d_{ic_1}, \dots, d_{ic_{n_i}})]$$

which is equal to

$$\begin{aligned} & \sum_{i=1}^n \sum_{d_f} \sum_{d_m} \sum_{d_{c_1}} \dots \sum_{d_{c_{n_i}}} [\log P(g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}}, d_f, d_m, d_{c_1}, \dots, d_{c_{n_i}})] \\ & \quad \times P(d_f, d_m, d_{c_1}, \dots, d_{c_{n_i}} | g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}}, p^{(j)}) \\ & = \sum_{i=1}^n \sum_{d_f} \sum_{d_m} \sum_{d_{c_1}} \dots \sum_{d_{c_{n_i}}} [\log \Phi_i + \log P(d_f, d_m) + \log P(d_{c_1}, \dots, d_{c_{n_i}} | d_f, d_m)] \cdot \Psi_i, \quad (7) \end{aligned}$$

where \sum_{d_f} and \sum_{d_m} denote summations over all possible diplotypes, $\sum_{d_{c_v}}$ ($v = 1, \dots, n_i$) means summation over those diplotypes which are Mendelian consistent with the diplotypes of parents, and Ψ_i is the posterior probability of diplotypes given the observed genotypes of the i th family:

$$\begin{aligned} \Psi_i & \equiv P(d_f, d_m, d_{c_1}, \dots, d_{c_{n_i}} | g_{if}, g_{im}, g_{ic_1}, \dots, g_{ic_{n_i}}, p^{(j)}) \\ & = \frac{\Phi_i P(d_f | p^{(j)}) P(d_m | p^{(j)}) \prod_{v=1}^{n_i} P(d_{c_v} | d_f, d_m)}{\sum_{d_f} \sum_{d_m} \sum_{d_{c_1}} \dots \sum_{d_{c_{n_i}}} \Phi_i P(d_f | p^{(j)}) P(d_m | p^{(j)}) \prod_{v=1}^{n_i} P(d_{c_v} | d_f, d_m)}. \end{aligned}$$

(ii) **M-step:**

We need to maximize $Q(p | p^{(j)})$ given by formula (7) with respect to p . Note that given the diplotypes of the parents, the diplotypes of their children do not depend on the parameter $p = (p_1, \dots, p_H)$, and

$$\frac{\partial \log P(d_f, d_m)}{\partial p_t} = \frac{\delta_{d_f, t} + \delta_{d_m, t}}{p_t},$$

where $\delta_{d_f, t}$ and $\delta_{d_m, t}$ denote the numbers of times that haplotype h_t is present in diplotype d_f and d_m , respectively. Based on these and the fact that $\sum_{t=1}^H \delta_{d_f, t} = \sum_{t=1}^H \delta_{d_m, t} = 2$, we can obtain the $(j+1)$ th iteration parameter estimates as:

$$\begin{aligned}
p_t^{(j+1)} &= \frac{\sum_{i=1}^n \sum_{d_f} \sum_{d_m} \sum_{d_{c_1}} \cdots \sum_{d_{c_{n_i}}} (\delta_{d_f,t} + \delta_{d_m,t}) \cdot \Psi_i}{\sum_{t=1}^H \sum_{i=1}^n \sum_{d_f} \sum_{d_m} \sum_{d_{c_1}} \cdots \sum_{d_{c_{n_i}}} (\delta_{d_f,t} + \delta_{d_m,t}) \cdot \Psi_i} \\
&= \frac{1}{4n} \sum_{i=1}^n \sum_{d_f} \sum_{d_m} \sum_{d_{c_1}} \cdots \sum_{d_{c_{n_i}}} (\delta_{d_f,t} + \delta_{d_m,t}) \cdot \Psi_i.
\end{aligned}$$

Based on the estimates of haplotype frequencies, $\hat{p}_1, \dots, \hat{p}_H$, we can estimate the number of families with genotype $G^* = (g_f^*, g_m^*, g_{c_1}^*, \dots, g_{c_V}^*)$, where V is the number of children in the family, in the sample as:

$$\begin{aligned}
n_{G^*} &= n \cdot P(g_f^*, g_m^*, g_{c_1}^*, \dots, g_{c_V}^*) \\
&= n \cdot \sum_{d_f}^* \sum_{d_m}^* \sum_{d_{c_1}}^* \cdots \sum_{d_{c_V}}^* P(d_f, d_m, d_{c_1}, \dots, d_{c_V}) \\
&= n \cdot \sum_{d_f}^* \sum_{d_m}^* \sum_{d_{c_1}}^* \cdots \sum_{d_{c_V}}^* P(d_f) P(d_m) \cdot \prod_{v=1}^V P(d_{c_v} | d_f, d_m),
\end{aligned}$$

where $\sum_{d_f}^*$, $\sum_{d_m}^*$, and $\sum_{d_{c_v}}^*$ ($v = 1, \dots, V$) denote summations over all different diplotypes consistent with genotypes g_f^* , g_m^* and $g_{c_v}^*$ ($v = 1, \dots, V$), respectively.

Simulation study

To evaluate the performance of our proposed approach, we conduct simulation studies outlined below.

We let the set of all possible haplotypes consist of all of the vectors (i_1, \dots, i_k) ($i_1 = 1, \dots, I_1; \dots; i_k = 1, \dots, I_k$). For the samples consisting of unrelated individuals, we first generate the diplotypes of n individuals by randomly drawing n pairs of haplotypes according to a set of haplotype frequencies $p_{i_1 \dots i_k}$ ($i_1 = 1, \dots, I_1; \dots; i_k = 1, \dots, I_k$). From these diplotypes we can derive each individual's marker genotypes. We then introduce genotyping errors independently for each marker alleles under the error model given above where, for simplicity, we assume the same genotyping error rates across markers. These error-contaminated marker genotypes are analyzed by our proposed EM algorithm to estimate haplotype frequencies.

Similarly, for the samples consisting of nuclear families, we first generate the diplotypes of the two parents for all n families by randomly drawing $2n$ pairs of haplotypes according to the haplotype frequencies $p_{i_1 \dots i_k}$ ($i_1 = 1, \dots, I_1; \dots; i_k = 1, \dots, I_k$). Based on parental diplotypes, we simulate diplotypes of children by randomly assigning one of the two haplotypes in each parent to each child. These lead to the true marker genotypes for each member in the nuclear families. We then introduce genotyping errors independently into the marker alleles of each member of the families under the error model given before where the error rates are also assumed to be the same across markers to obtain the observed error-contaminated family genotypes. We finally use the EM algorithm described above to estimate haplotype frequencies.

Results

In our simulations, we consider two error rates $\varepsilon = 0.01$ and 0.05 , respectively. We conduct 100 simulations to estimate the means and standard deviations of haplotype frequency estimates.

For the case of unrelated individual data, we take the sample size to be $n = 200$. Tables 2 and 3 summarize the simulation results on the haplotype frequency estimates and their standard deviations for haplotypes consisting of two and three biallelic markers, respectively. It is clear that the precision of our estimates is satisfactory for the error rates usually encountered in practice.

For nuclear families, we take the sample size to be $n = 100$. Tables 4 and 5 summarize the simulation results on the haplotype frequency estimates and their standard deviations for two biallelic markers based on family trios and quartets, respectively. It can be seen that as in the case of unrelated individuals, for family data, the estimates of haplotype frequencies are also quite close to their true values in the presence of genotyping errors.

Comparing the results under the error rates of $\varepsilon = 0.01$ and 0.05 , we see that overall, the estimates of haplotype frequencies are equally close to their true values under different error rates. We can also note that the standard deviations are slightly

greater for the larger genotyping error rate.

Discussion

In this article, we have developed statistical methods to incorporate genotyping errors to estimate haplotype frequencies both for samples consisting of unrelated individuals and for samples consisting of nuclear families. Simulation results show that our methods perform well. From the haplotype frequency estimates, we can also estimate the number of individuals/families with true genotypes, and study the impact of genotyping errors on any existing statistical methods for analyzing genotype data under our error model. The computer program will be made available at our web site <http://bioinformatics.med.yale.edu>.

For the specific case of one biallelic marker and unrelated individuals, equation (1) suggests that the allele frequency estimates are monotonic in the error rate when the error rates from allele 1 to allele 2 and from allele 2 to allele 1, ε_1 and ε_2 , are assumed to be equal. Therefore, neglecting genotyping errors may lead to substantial bias in haplotype frequency estimates. For the more general case, we generate sample data assuming the presence of genotyping errors, and then estimate haplotype frequencies ignoring errors. We find that ignoring genotyping errors may result in large bias. For example, consider the case of unrelated individual data summarized in Table 2. When the true error rate is 0.05, and we neglect the genotyping errors, then the estimates based on the averages of 100 simulations (and their standard deviations) are 0.366 (0.0025), 0.126 (0.0018), 0.132 (0.0018) and 0.376 (0.0025), respectively. Therefore, the relative bias can be as high as 30% when the genotyping error rate is 5%.

Note that the error rates are assumed to be known in our methods here. If the error rates are unknown, we may estimate them using different approaches. One simple approach is to genotype a set of samples with known genotypes multiple times and estimate genotyping error rates from these multiple measurements. Then the estimated error rate can be used for the same genotyping technique when we analyze

a set of data with unknown marker genotypes. Alternatively, we may simultaneously estimate haplotype frequencies and genotyping error rates if the samples are genotyped more than once. For example, in the case of single biallelic markers and unrelated individuals, there are nine possible observed genotype combinations for an individual if everyone is genotyped twice: $(1\ 1)' \times (1\ 1)'$, $(1\ 1)' \times (1\ 2)'$, $(1\ 1)' \times (2\ 2)'$, $(1\ 2)' \times (1\ 1)'$, $(1\ 2)' \times (1\ 2)'$, $(1\ 2)' \times (2\ 2)'$, $(2\ 2)' \times (1\ 1)'$, $(2\ 2)' \times (1\ 2)'$ and $(2\ 2)' \times (2\ 2)'$ with respective probabilities $[(1 - \varepsilon)^2 p + \varepsilon^2 q]^2$, $2\varepsilon(1 - \varepsilon)[(1 - \varepsilon)^2 p + \varepsilon^2 q]$, $\varepsilon^2(1 - \varepsilon)^2$, $2\varepsilon(1 - \varepsilon)[(1 - \varepsilon)^2 p + \varepsilon^2 q]$, $4\varepsilon^2(1 - \varepsilon)^2 + 2pq(1 - 2\varepsilon)^2$, $2\varepsilon(1 - \varepsilon)[\varepsilon^2 p + (1 - \varepsilon)^2 q]$, $\varepsilon^2(1 - \varepsilon)^2$, $2\varepsilon(1 - \varepsilon)[\varepsilon^2 p + (1 - \varepsilon)^2 q]$, and $[\varepsilon^2 p + (1 - \varepsilon)^2 q]^2$, where ε is the error rate from one allele to the other allele, p is the frequency of allele 1 and $q = 1 - p$. If we denote the numbers of their observations by n_{11} , n_{12} , n_{13} , n_{21} , n_{22} , n_{23} , n_{31} , n_{32} , and n_{33} , respectively, then the log-likelihood is given by (neglecting constants)

$$\begin{aligned}
& (2n_{11} + n_{12} + n_{21}) \cdot \log [(1 - \varepsilon)^2 p + \varepsilon^2 q] + (n_{23} + n_{32} + 2n_{33}) \cdot \log [\varepsilon^2 p + (1 - \varepsilon)^2 q] \\
& + [n_{12} + n_{21} + 2(n_{13} + n_{31}) + n_{23} + n_{32}] \cdot \log [\varepsilon(1 - \varepsilon)] \\
& + n_{22} \cdot \log [2\varepsilon^2(1 - \varepsilon)^2 + pq(1 - 2\varepsilon)^2]. \tag{9}
\end{aligned}$$

By maximizing this log-likelihood function simultaneously for p and ε , we can obtain the estimates for both p and ε . For example, consider the case of true allele frequency $p = 0.9$ and true error rate $\varepsilon = 0.01$. For the sample of size 1000, we assume that the numbers of observations for the nine genotype combinations are 778, 17, 0, 17, 174, 2, 0, 2 and 10, respectively (Note that these are just the expected numbers of the nine genotype combinations in the sample except that of n_{22} should be 173). Then the log-likelihood (9) does attain its maximum *only* at $(p, \varepsilon) = (0.899, 0.010)$. If we do genotyping only once and we assume that the numbers of observations for the genotypes $(1\ 1)'$, $(1\ 2)'$ and $(2\ 2)'$ are 810, 180 and 10, respectively, then the corresponding log-likelihood can attain its maximum for those pairs of (p, ε) satisfying $p(1 - 2\varepsilon) = 0.9 - \varepsilon$.

For the case of family data, we have assumed that the parental genotype data is available. The treatments for the other cases where the genotypes of one or both

parents are not available or pedigree data is available should be not difficult using the estimation framework discussed in this article.

Acknowledgments

This work was supported in part by grant GM59507 from the National Institutes of Health. We thank two reviewers for their valuable comments and suggestions.

References

- Akey JM, Zhang K, Xiong M, Doris P, and Jin L (2001). The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am J Hum Genet* 68: 1447-1456.
- Broman KW, Weber JL (1998). Estimation of pairwise relationships in the presence of genotyping errors. *Am J Hum Genet* 63: 1563-1564.
- Buetow KH (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am J Hum Genet* 49: 985-994.
- Clark AG (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7: 111-122.
- Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via EM algorithm. *J R Stat Soc Ser B* 39: 1-38.
- Excoffier L, Slatkin M (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12: 921-927.
- Excoffier L, Slatkin M (1998). Incorporating genotypes of relatives into a test of linkage disequilibrium. *Am J Hum Genet* 62: 171-180.
- Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, and Schork NJ (2001). Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and Alzheimer's disease. *Genome Res* 11: 143-151.
- Goldstein DR, Zhao H, and Speed TP (1997). The effects of genotyping errors and interference on estimation of genetic distance. *Hum Hered* 47: 86-100.
- Gordon D, Heath SC, Liu X, and Ott J (2001). A transmission /disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* 69: 371-380.
- Gordon D, Heath SC, and Ott J (1999). True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum Hered* 49: 65-70.
- Gordon D, Ott J (2001). Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac Symp Biocomput*

2001: 18-29.

Göring HHH, Terwilliger JD (2000*a*). Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet* 66: 1095-1106.

Göring HHH, Terwilliger JD (2000*b*). Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am J Hum Genet* 66: 1107-1118.

Göring HHH, Terwilliger JD (2000*c*). Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. *Am J Hum Genet* 66: 1298-1309.

Göring HHH, Terwilliger JD (2000*d*). Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* 66: 1310-1327.

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

Long JC, Williams RC, and Urbanek M (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 56: 799-810.

Niu T, Qin Z, Xu X., and Liu JS (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 70: 157-169.

Schaid DJ (2002). Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genet Epidemiol* 23: 426-443.

Shields DC, Collins A, Buetow KH, and Morton NE (1991). Error filtration, interference, and the human linkage map. *Proc Natl Acad Sci USA* 88: 6501-6505.

Sobel E, Papp J, and Lange K (2002). Detection and Integration of genotyping errors in statistical genetics. *Am J Hum Genet* 70: 496-508.

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

Terwilliger JD, Weeks DE, and Ott J (1990). Laboratory errors in the reading of

marker alleles cause massive reductions in lod score and lead to gross overestimates of the recombination fraction. *Am J Hum Genet Suppl* 47: A201.

Zou G, Pan D, and Zhao H (2003). Genotyping error detection through tightly linked markers. *Genetics*, in press.

Table 2. Estimation of haplotype frequencies
for $k = 2$, $I_1 = I_2 = 2$ and $n = 200$ based on unrelated individuals

haplotype	true frequency	estimated frequency (standard deviation)	
		$\varepsilon = 0.01$	$\varepsilon = 0.05$
(1 1)	0.4	0.396 (0.0026)	0.392 (0.0029)
(1 2)	0.1	0.099 (0.0016)	0.099 (0.0021)
(2 1)	0.1	0.102 (0.0016)	0.104 (0.0022)
(2 2)	0.4	0.403 (0.0024)	0.405 (0.0027)

Table 3. Estimation of haplotype frequencies
for $k = 3$, $I_1 = I_2 = I_3 = 2$ and $n = 200$ based on unrelated individuals

haplotype	true frequency	estimated frequency (standard deviation)	
		$\varepsilon = 0.01$	$\varepsilon = 0.05$
(1 1 1)	0.4	0.399 (0.0028)	0.400 (0.0036)
(1 1 2)	0.2	0.202 (0.0026)	0.204 (0.0031)
(1 2 1)	0.1	0.102 (0.0020)	0.101 (0.0025)
(1 2 2)	0.1	0.097 (0.0020)	0.094 (0.0023)
(2 1 1)	0.08	0.081 (0.0017)	0.082 (0.0021)
(2 1 2)	0.05	0.048 (0.0015)	0.049 (0.0016)
(2 2 1)	0.05	0.046 (0.0013)	0.047 (0.0018)
(2 2 2)	0.02	0.025 (0.0012)	0.024 (0.0013)

Table 4. Estimation of haplotype frequencies
for $k = 2$, $I_1 = I_2 = 2$ and $n = 100$ based on family trios

haplotype	true frequency	estimated frequency (standard deviation)	
		$\varepsilon = 0.01$	$\varepsilon = 0.05$
(1 1)	0.4	0.395 (0.0025)	0.393 (0.0029)
(1 2)	0.1	0.100 (0.0016)	0.102 (0.0021)
(2 1)	0.1	0.102 (0.0016)	0.104 (0.0018)
(2 2)	0.4	0.403 (0.0025)	0.402 (0.0029)

Table 5. Estimation of haplotype frequencies
for $k = 2$, $I_1 = I_2 = 2$ and $n = 100$ based on family quartets

haplotype	true frequency	estimated frequency (standard deviation)	
		$\varepsilon = 0.01$	$\varepsilon = 0.05$
(1 1)	0.4	0.403 (0.0027)	0.402 (0.0030)
(1 2)	0.1	0.102 (0.0016)	0.102 (0.0018)
(2 1)	0.1	0.101 (0.0015)	0.102 (0.0017)
(2 2)	0.4	0.394 (0.0024)	0.394 (0.0027)