Biostatistics (2003), **4**, 4, *pp.* 513–522
Printed in Great Britain

Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms

SHUYING SUE LI, NAJMA KHALID
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
CHRISTOPHER CARLSON
Department of Genome Sciences, University of Washington, Seattle, WA, USA
LUE PING ZHAO*
Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue. N., Seattle, WA 98109, USA
lzhao@fhcrc.org

SUMMARY

Estimating haplotype frequencies becomes increasingly important in the mapping of complex disease genes, as millions of single nucleotide polymorphisms (SNPs) are being identified and genotyped. When genotypes at multiple SNP loci are gathered from unrelated individuals, haplotype frequencies can be accurately estimated using expectation-maximization (EM) algorithms (Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long *et al.*, 1995), with standard errors estimated using bootstraps. However, because the number of possible haplotypes increases exponentially with the number of SNPs, handling data with a large number of SNPs poses a computational challenge for the EM methods and for other haplotype inference methods. To solve this problem, Niu and colleagues, in their Bayesian haplotype inference paper (Niu *et al.*, 2002), introduced a computational algorithm called progressive ligation (PL). But their Bayesian method has a limitation on the number of subjects (no more than 100 subjects in the current implementation of the method). In this paper, we propose a new method in which we use the same likelihood formulation as in Excoffier and Slatkin's EM algorithm and apply the estimating equation idea and the PL computational algorithm with some modifications. Our proposed method can handle data sets with large number of SNPs as well as large numbers of subjects. Simultaneously, our method estimates standard errors efficiently, using the sandwich-estimate from the estimating equation, rather than the bootstrap method. Additionally, our method admits missing data and produces valid estimates of parameters and their standard errors under the assumption that the missing genotypes are *missing at random* in the sense defined by Rubin (1976).

Keywords: Estimating equation; Haplotype; Hardy–Weinberg equilibrium; Single nucleotide polymorphism (SNP).

1. INTRODUCTION

Genetic variation in human individuals can mostly be attributed to alterations in genomic DNA sequences (Wang *et al.*, 1998; Cargill *et al.*, 1999; Halushka *et al.*, 1999). Alterations that involve a single

*To whom correspondence should be addressed.

base pair and are observed in at least 5% of the population are called single nucleotide polymorphisms (SNPs). It is expected that the human genome contains millions of SNPs (Kruglyak and Nickerson, 2001). The challenge for statisticians is how to use the high-dimension SNP data efficiently to study human evolutionary history and human complex diseases such as diabetes, hypertension and cancer. Recent empirical evidence shows that these SNPs are divided into several blocks, with each block highly structured into a small number of haplotypes (Reich *et al.*, 2001; Daly *et al.*, 2001; Goldstein, 2001; Patil *et al.*, 2001). (A haplotype is defined as a sequence of alleles from the same chromosome. Each individual has two haplotypes because each individual has two copies of each chromosome inherited from his/her parents.) Constructing haplotypes from genotypes at multiple loci thus serves as a natural data reduction tool.

However, directly constructing haplotypes can be very difficult. Several experimental approaches, such as dissecting a single chromosome or inserting an entire chromosome into a yeast artificial chromosome (Green *et al.*, 1998) or using rodent–human hybrid techniques to physically separate two chromosomes (Patil *et al.*, 2001), have attempted to read alleles from each separated chromosome, but the techniques remain extremely expensive for wide scale use. SNP data are typically obtained in the form of genotypes at separate loci, without knowing the parental origins of alleles (see an example in Table 1). In order to gain some information about the parental origins of alleles in the genotypes, one method is to obtain the parental genotypes as well (Wijsman, 1987). However, collecting parental biological samples for genotyping not only increases the cost but may sometimes be impossible. As an alternative, statistical methods have been used to make inferences of individuals' haplotypes from their genotypes without requiring information about parental genotypes. Popular methods include the maximum likelihood approach using expectation-maximization (EM) algorithms (Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long *et al.*, 1995). Correct inference for individuals' haplotypes relies entirely on a correct estimation of haplotype frequencies in the population. Only accurate estimation of the haplotype frequencies assures an accurate inference for individuals' haplotypes from their genotypes. Thus, the haplotype inference methods are evaluated by the error measure on the estimates of haplotype frequencies. Fallin and Schork (2000) have shown that the EM algorithm (Excoffier and Slatkin, 1995) produces accurate estimates of haplotype frequencies for a wide range of parameter settings. More recently, two new methods using a Bayesian approach (Stephens *et al.*, 2001; Niu *et al.*, 2002) were proposed based on the likelihood formulation used in Excoffier and Slatkin (1995). Stephens *et al.* (2001) further assumed that the distribution of haplotypes satisfies coalescence theory. Simulation studies (Stephens *et al.*, 2001; Niu *et al.*, 2002) showed that the performances of the EM algorithm and Niu *et al.*'s Bayesian method on the prediction of individuals' haplotypes are similar to each other and are both generally better than the performance of Stephens *et al.*'s Bayesian method except when the data satisfy the additional assumptions required by the Bayesian method.

Handling data with a large number of SNPs poses a computational challenge for any haplotype inference method, because the number of possible haplotypes increases exponentially in the number of SNPs. The computational algorithm (progressive ligation (PL)) introduced in Niu *et al.* (2002) helps to solve part of the problem. But handling data with a large number of subjects appears to be an additional challenge for the Bayesian method. To overcome these challenges, we propose a new approach based on the same likelihood formulation used in Excoffier and Slatkin (1995) but adopting the estimating equation idea along with a modified PL computational algorithm. We consider our approach an improvement on the EM algorithm in three aspects: (1) our method is able to handle much larger data sets, both in terms of the number of SNPs and the number of subjects; (2) our algorithm is computationally efficient since we calculate the standard errors analytically without requiring a computationally intensive method such as bootstrap and (3) we are able to incorporate missing genotypes. The paper is organized as follows. In Section 2, we introduce the methodology. In Section 3, we describe simulation studies conducted to assess the statistical properties of the estimates and the computational efficiency of the proposed method.

Table 1. *Part of genotype data for 44 unrelated individuals with 18 SNPs within gene ARHGDIB on chromosome 12 (Reich* et al., *2001). Genotype 0: homozygous reference genotype; 1: heterozygous genotype; 2: homozygous variant genotype; 3: missing data. The first allele is used as reference in this example*

SNP's ID	Alleles	Individual's genotypes
G4923a44	T/C	00111112022121011100101101210010000001210012
G4923a46	C/T	00111112022121011100101101210010000001210012
G4923a35	C/T	00110012022121011100101101210011000001210012
G4923a37	A/G	10110121000100101002000010121103200001211110
G4923a38	A/G	10000000022001001013102301131020000001000000

In Section 4, we illustrate our proposed method using the data shown in Table 1 and compare our estimates with estimates using Arlequin (a software implementing the Excoffier and Slatkin's EM algorithm), and compare the computational efficiency of our method with other haplotype inference methods.

2. METHODS

Suppose we have a sample of n unrelated individuals from a population. From each individual, we observe q SNP-genotypes on a specific region in the genome, e.g. on a candidate gene. Let $\underset{\sim}{g}_i = (g_{i1}, g_{i2}, \ldots, g_{iq})$ denote the q SNP-genotypes for the ith individual where $g_{ij} = 0, 1, 2$ denotes homozygous reference, heterozygous and homozygous variant genotype, respectively. When g_{ij} is homozygous (0 or 2), the phase of genotype g_{ij} is unambiguous; when g_{ij} is heterozygous (1), the phase of genotype g_{ij} becomes ambiguous with two possible resolutions. Let p_{ij} denote the phase of g_{ij} and $\underset{\sim}{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iq})$ denote the phase of $\underset{\sim}{g}_i$. Because we do not try to identify the parental origin of each haplotype, $\underset{\sim}{p}_i$ of $\underset{\sim}{g}_i$ with m heterozygous genotypes have 2^{m-1} possible resolutions that give rise to 2^{m-1} possible pairs of haplotypes. Given each resolution of $\underset{\sim}{p}_i$, then $\underset{\sim}{g}_i$ is formed by a pair of haplotypes denoted by $\underset{\sim}{H}_i^1$ and $\underset{\sim}{H}_i^2$. For example, in the genotype data given in Table 1, the first individual has genotype $\underset{\sim}{g}_1 = (00011)$ with two heterozygous loci, therefore there are two possible resolutions for phase $\underset{\sim}{p}_1$. Haplotypes $H_1^1 = (00000)$ and $H_1^2 = (00011)$, where allele 0 represents a reference allele and 1 represents a variant allele, result from one of resolutions of $\underset{\sim}{p}_1$.

For q SNP loci there are $T = 2^q$ possible haplotypes. Let $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_T)$ denote the unknown haplotype frequencies with $\sum_{i=1}^{T} \theta_i = 1$. The distribution of haplotypes is assumed to be multinomial with parameter $\underset{\sim}{\theta}$. Under the assumption of the Hardy–Weinberg Equilibrium (HWE), the conditional distribution of $\underset{\sim}{g}_i$ given $\underset{\sim}{p}_i$ is a product of the frequencies of two associated haplotypes and the likelihood function of $\underset{\sim}{\theta}$ can be expressed as

$$L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} f(\underset{\sim}{g}_i; \underset{\sim}{\theta}) = \prod_{i=1}^{n} \sum_{\underset{\sim}{p}_i} f(\underset{\sim}{g}_i | \underset{\sim}{p}_i; \underset{\sim}{\theta}) f(\underset{\sim}{p}_i) = \prod_{i=1}^{n} 2^{-c_i} \sum_{\underset{\sim}{p}_i : \underset{\sim}{g}_i | \underset{\sim}{p}_i = (\underset{\sim}{H}_i^1, \underset{\sim}{H}_i^2)} \theta_{\underset{\sim}{H}_i^1} \theta_{\underset{\sim}{H}_i^2}, \qquad (2.1)$$

where $f(\underset{\sim}{p}_i) = 2^{-c_i}$ and c_i is the number of heterozygous loci minus one. After algebraic manipulations,

the estimating equations derived from the likelihood function are expressed as

$$\underset{\sim}{U}(\underset{\sim}{\theta}) = \sum_{i=1}^{n} U_i(\underset{\sim}{\theta}) = \sum_{i=1}^{n} E_{\underset{\sim}{p}_i}(\underset{\sim}{F}_i \mid \underset{\sim}{g}_i; \underset{\sim}{\theta}) = \sum_{i=1}^{n} \sum_{\underset{\sim}{p}_i} \underset{\sim}{F}_i \, f(\underset{\sim}{p}_i \mid \underset{\sim}{g}_i; \underset{\sim}{\theta}) = \underset{\sim}{0}, \tag{2.2}$$

where $\underset{\sim}{F}_i = (F_{i1}, F_{i2}, \ldots, F_{iT})'$ in which $F_{ij} = I(\underset{\sim}{H}_i^1 = \underset{\sim}{h}_j) + I(\underset{\sim}{H}_i^2 = \underset{\sim}{h}_j) - 2\theta_j$ is the difference between the observed and the expected frequencies of the jth haplotypes from the ith individual, $I(\cdot)$ is an indicator function and

$$f(\underset{\sim}{p}_i \mid \underset{\sim}{g}_i; \underset{\sim}{\theta}) = \frac{f(\underset{\sim}{g}_i \mid \underset{\sim}{p}_i; \underset{\sim}{\theta}) f(\underset{\sim}{p}_i)}{\sum_{\underset{\sim}{p}_i} f(\underset{\sim}{g}_i \mid \underset{\sim}{p}_i; \underset{\sim}{\theta}) f(\underset{\sim}{p}_i)} = \frac{\theta_{\underset{\sim}{H}_i^1} \theta_{\underset{\sim}{H}_i^2}}{\sum_{\underset{\sim}{p}_i : \underset{\sim}{g}_i \mid \underset{\sim}{p}_i = (\underset{\sim}{H}_i^1, \underset{\sim}{H}_i^2)} \theta_{\underset{\sim}{H}_i^1} \theta_{\underset{\sim}{H}_i^2}} \tag{2.3}$$

is the posterior distribution of the phase $\underset{\sim}{p}_i$ given genotype $\underset{\sim}{g}_i$ (details of the derivations of the likelihood and the estimating equations are given in the technical report in our website `http://qge.fhcrc.org/hplus/`).

Despite advances in genotyping technologies, a small fraction of genotypes (usually less than 1%) cannot be determined due to technical reasons, and are treated as missing values. There are three possibilities for missing genotypes: missing both alleles ($g_{ij} = 3$) and missing one allele while the other observed allele is the reference allele ($g_{ij} = 4$) or the variant allele ($g_{ij} = 5$). Let $\underset{\sim}{g}_i = (\underset{\sim}{g}_i^O, \underset{\sim}{g}_i^M)$ where superscripts O and M represent observed and missing genotypes, respectively, for the ith individual. Then, the estimating equations (2.2) can be modified as follows:

$$\underset{\sim}{U}(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \sum_{\underset{\sim}{g}_i^M} \sum_{\underset{\sim}{p}_i \mid \underset{\sim}{g}_i} \underset{\sim}{F}_i \, f(\underset{\sim}{p}_i \mid \underset{\sim}{g}_i; \underset{\sim}{\theta}) \, f(\underset{\sim}{g}_i^M \mid \underset{\sim}{g}_i^O; \underset{\sim}{\theta}) = \underset{\sim}{0} \tag{2.4}$$

where the conditional distribution of missing genotypes given observed genotypes, $f(\underset{\sim}{g}_i^M \mid \underset{\sim}{g}_i^O; \underset{\sim}{\theta})$, is specified based on the missing mechanism. Without prior knowledge of the missing mechanism, we assume that the missing genotypes are *missing at random* (MAR) in the sense defined by Rubin (1976) and that the distribution of the missing genotypes depends upon the observed genotypes and the distribution of haplotypes, i.e.

$$f(\underset{\sim}{g}_i^M \mid \underset{\sim}{g}_i^O; \underset{\sim}{\theta}) = \frac{\sum_{\underset{\sim}{p}_i \mid \underset{\sim}{g}_i} f(\underset{\sim}{p}_i \mid \underset{\sim}{g}_i = (\underset{\sim}{g}_i^O, \underset{\sim}{g}_i^M); \underset{\sim}{\theta})}{\sum_{\underset{\sim}{g}_i^M} \sum_{\underset{\sim}{p}_i \mid \underset{\sim}{g}_i} f(\underset{\sim}{p}_i \mid \underset{\sim}{g}_i = (\underset{\sim}{g}_i^O, \underset{\sim}{g}_i^M); \underset{\sim}{\theta})}.$$

We can find the estimate of $\underset{\sim}{\theta}$ by iteratively evaluating $\underset{\sim}{\theta}$ using (2.2) if there are no missing genotypes in the data or from (2.4) if there are some missing genotypes, where $f(\underset{\sim}{p}_i \mid \underset{\sim}{g}_i; \underset{\sim}{\theta})$ is evaluated using (2.3) for the solution of $\underset{\sim}{\theta}$ in the previous iteration, until convergence is reached. In the first iteration, we set $f(\underset{\sim}{p}_i \mid \underset{\sim}{g}_i; \underset{\sim}{\theta}) = f(\underset{\sim}{p}_i) = 2^{-c_i}$ assuming that all phases have an equal probability.

Consequently, any haplotypes not observed in the first iteration must have the final estimates of their frequencies equal to zero. This estimation strategy usually works for data sets with a small number of SNPs because the number of possible resolutions of phase increases exponentially with the number of heterozygous loci. For a large data set, we use an algorithm modified from the progressive ligation

computational algorithm (Niu *et al.*, 2002). Briefly, the entire set of SNPs is divided into several small blocks with K SNPs (e.g. $K = 10$). The estimation is first performed separately for the first two blocks and then performed for the joined block. This estimation process is repeated for the joined block and the next single block until the last block has been joined. The estimation for each single block is done exactly as described above. The estimation for each joined block is done similarly except that estimation results from the previous steps are used as initial values. After each of estimation, haplotypes with estimated frequencies above δ are retained in the following estimation process. δ is chosen to be small enough to ensure that no final haplotypes with nonzero frequencies will be discarded in an intermediate step (e.g. $\delta = 10^{-5}$). This is different from Niu *et al.* (2002) where B most probable haplotypes are retained.

The estimates of haplotype frequencies in the final step correspond to the estimates for the entire data. Suppose that there are m haplotypes whose estimated frequencies are nonzero in the final estimation step. The covariance of the first $m - 1$ nonzero elements in $\hat{\underset{\sim}{\theta}}$ is estimated by

$$\Sigma(\hat{\underset{\sim}{\theta}}) = (\partial U(\hat{\underset{\sim}{\theta}})/\partial\hat{\underset{\sim}{\theta}})^{-1}\text{var}[U(\hat{\underset{\sim}{\theta}})](\partial U(\hat{\underset{\sim}{\theta}})/\partial\hat{\underset{\sim}{\theta}})^{-1} \tag{2.5}$$

with $\frac{\partial U(\hat{\underset{\sim}{\theta}})}{\partial\hat{\underset{\sim}{\theta}}} = -\sum_i[2I - \text{var}(\underset{\sim}{F}_i|\underset{\sim}{g}_i;\hat{\underset{\sim}{\theta}})V^{-1}]$, where V is the covariance matrix of a multinomial distribution with diagonal elements equal to $\hat{\theta}_i(1-\hat{\theta}_i)$ and off-diagonal elements equal to $-\hat{\theta}_i\hat{\theta}_j$, and $\text{var}[U(\hat{\underset{\sim}{\theta}})] = \sum_i E_{\underset{\sim}{p}_i}(\underset{\sim}{F}_i|\underset{\sim}{g}_i;\hat{\underset{\sim}{\theta}})E_{\underset{\sim}{p}_i}(\underset{\sim}{F}_i'|\underset{\sim}{g}_i;\hat{\underset{\sim}{\theta}})$. The standard error of $\hat{\underset{\sim}{\theta}}$ is estimated by the square root of diagonals of $\Sigma(\hat{\underset{\sim}{\theta}})$. The standard error of $\hat{\theta}_m$, the estimated frequency of the least frequency haplotype, is estimated based the condition $\sum_{i=1}^m \hat{\theta}_i = 1$ and $\Sigma(\hat{\underset{\sim}{\theta}})$. However, if $\hat{\underset{\sim}{\theta}}$ includes some rare haplotypes, the calculation of covariance in (2.5) may not be stable. One solution is to collapse all rare haplotypes into one composite haplotype and to estimate the standard error of the estimated frequency of the composite haplotype. A haplotype is considered to be rare here if its estimated haplotype frequency satisfies the conditions $2n\hat{\theta}_j < 5$ and $\hat{\theta}_j < 0.01$.

According to estimating equation theory Zeger and Liang (1986); Zhao *et al.* (1998), the asymptotic distribution of $\hat{\underset{\sim}{\theta}}$ is normal with mean $\underset{\sim}{\theta}$ and covariance $\Sigma(\underset{\sim}{\theta})$ provided that the conditional probability $f(\underset{\sim}{p}_i|\underset{\sim}{g};\underset{\sim}{\theta})$ is correctly specified. Thus, the confidence interval of estimate of $\hat{\theta}_j$ is estimated by $(\hat{\theta}_j - Z_{(1-\alpha)}\hat{\sigma}_{\hat{\theta}_j}, \hat{\theta}_j + Z_{(1-\alpha)}\hat{\sigma}_{\hat{\theta}_j})$.

3. Simulation study

It has been shown that under the assumption of HWE, the estimates of haplotype frequencies and their SE derived from estimating equations (2.2) and (2.5), theoretically approach to their 'true' values as the sample size goes to infinity (Liang and Zeger, 1986). For finite sample sizes, using simulations, we assess these statistical properties of the estimates both under the model assumption and when the model assumption is violated. We also assess the scalability of our proposed method to large numbers of SNPs. Each simulation is done in three steps: (1) generating a distribution of population haplotypes using a coalescence-based program (Hudson, 2002) to simulate the population evolutionary process; (2) generating each individual's genotype by sampling a pair of haplotypes from the distribution of population haplotypes and (3) estimating haplotype frequencies and standard errors (SE). For each generated distribution in the first step, we repeat steps 2 and 3 for 500 times. We then repeat the entire process 20 times.

To assess the accuracy of the estimated haplotype frequencies, we compare them to the actual haplotype frequencies, either in the population or in the observed sample. We use two similarity indices, $I_G = 1 - 0.5\sum_j |\hat{\theta}_j - \theta_j|$ and $I_S = 1 - 0.5\sum_j |\hat{\theta}_j - \tilde{\theta}_j|$, where $\theta_j, \hat{\theta}_j$, and $\tilde{\theta}_j$ denote the actual and estimated haplotype frequencies in the population and actual haplotype frequencies in the observed sample, respectively, and $\sum_j$ is the summation over all haplotypes. The similarity index I_G (or I_S) varies between zero and one. The two indices assess different aspects of the estimation. The index I_G assesses the overall validity of the final estimates of haplotype frequencies with respect to the true population values. The index I_S assesses the accuracy of the estimation method. To assess the accuracy of the estimated standard errors $\hat{\underset{\sim}{\theta}}$ denoted by the vector $\hat{\underset{\sim}{\sigma}}$, over all common haplotypes, we compare them to the sample standard deviations of the estimates, denoted by the vector $\tilde{\underset{\sim}{\sigma}}$, using the average mean square error, avg(MSE) $= \frac{1}{c}\sum_{j=1}^{c}(\hat{\sigma}_j - \tilde{\sigma}_j)^2$, over all haplotypes with $2n\hat{\theta}_j \geqslant 5$ or $\hat{\theta}_j \geqslant 0.01$ and the composite haplotype of the remaining (rare) haplotypes.

In the first set of simulations, we assess the accuracy of the estimates of haplotype frequencies and standard errors under the assumption of HWE and when HWE is violated, called Hardy–Weinberg disequilibrium (HWD). Under HWE, individuals' genotypes are generated by randomly sampling a pair of haplotypes from the population. Under HWD, two scenarios are considered: (1) one locus carries a lethal mutant allele such that individuals with a heterozygous genotype (0/1) at that locus have a 50% chance for survival and individuals with a homozygous variant genotype (1/1) have no chance for survival and (2) one haplotype is associated with a disease such that individuals with one copy of that haplotype have a 75% chance for survival and individuals with two copies have a 50% chance for survival. In each scenario, individuals' genotypes are generated accordingly. The data are generated for sample sizes of 30, 50, 100, 150, 200, 500 and 1000, with 10 SNPs and a recombination rate $R = 4$ (a parameter used in the coalescence-based program to describe evolutionary processes in the population, different from the recombination fraction, a measure of distance between two genetic markers). The simulation results are shown in Figure 1. The upper panel presents the two similarity indices I_G (dashed line) and I_S (solid line), under the assumption of HWE, HWD(1), and HWD(2), respectively. The index I_G is always smaller than the index I_S because of sampling error. The index I_S is above 0.9 even for a sample size of 30 and approaches one for larger sample sizes. This result indicates that our proposed method works very well even for the smallest sample size of 30. However, I_G falls below 0.9 when the sample size is less than 50 but increases rapidly as the sample size increases, and approaches the index I_S. Neither index is significantly affected by departures from the HWE. The lower panel in Figure 1 shows the average MSE of the estimated SE $\hat{\sigma}_j$, compared to the sample SD of the estimates $\tilde{\sigma}_j$. The average MSE is small, (below 2.5×10^{-4}), even for the lowest sample size of 30 and drops sharply as sample sizes increase from 30 to 100. But, beyond a sample size of 100, there is a much more gradual decrease in the average MSE. The average MSE is affected by the departures from the HWE only when sample size is small. In this set of simulations, the average number of haplotypes was 11. The average number of haplotypes with $2n\hat{\theta}_j \geqslant 5$ or $\hat{\theta}_j \geqslant 0.01$ ranged from 5 to 9.

In the second set of simulations, we assess accuracy of estimates of haplotype frequencies and their SE for a finite sample size and large numbers of SNPs. The data are generated under the assumption of HWE with a sample size of 100 and the number of SNPs to be 20, 40, 60, 80 and 100 for two different recombination rates, $R = 0$, 4, and 40. The simulation results are shown in Figure 2. For a sample size of 100, I_G is above 0.9 when the number of SNPs is less than 80 and there is no recombination ($R = 0$) or a small rate of recombination. Moreover, the average MSE of the estimated SE of the estimates is consistently small for all selected numbers of SNPs and recombination rates. The average numbers of observed population haplotypes ranged from 17 to 59 for $R = 0$, from 21 to 65 for $R = 4$, and 42 to 135 for $R = 40$. The average number of haplotypes that satisfied either $2n\hat{\theta}_j \geqslant 5$ or $\hat{\theta}_j \geqslant 0.01$, ranged from 6 to 12.

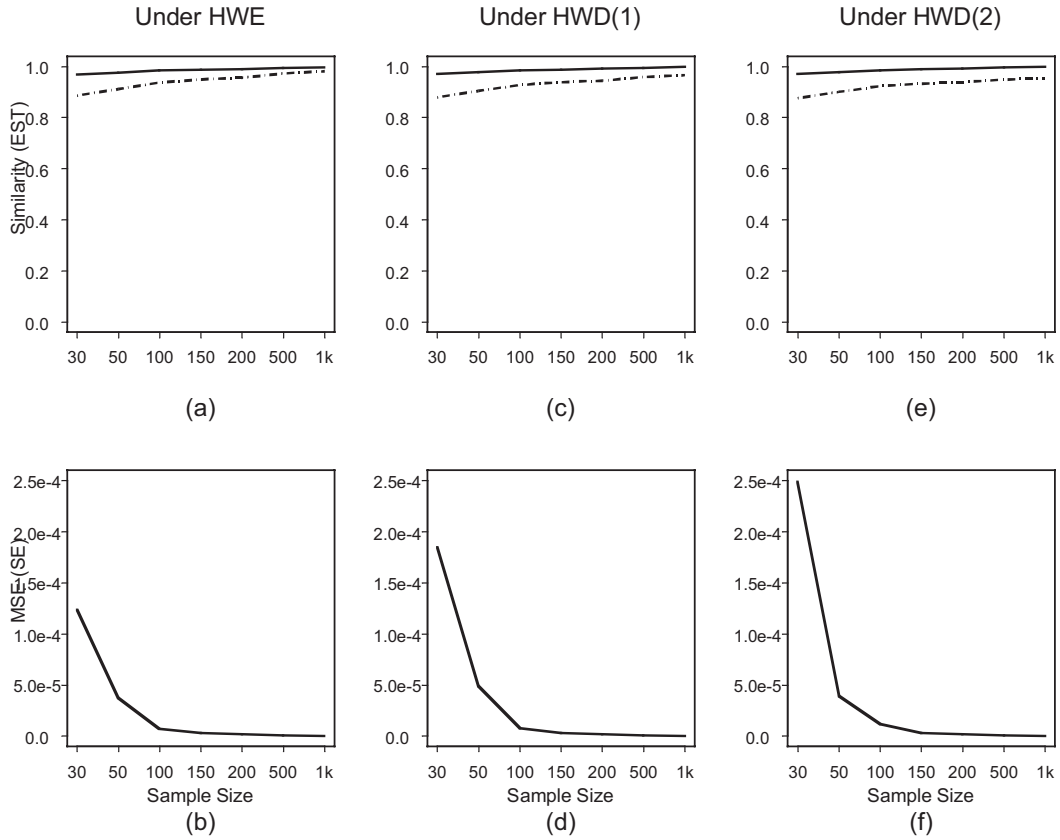

Fig. 1. The upper panel shows the similarity index I_G between the estimates using our proposed method and the actual haplotype frequencies in the population (dashed line) and the similarity index I_S between the estimates and the actual haplotype frequencies in the observed sample (solid line), under HWE, HWD(1) and HWD(2), respectively. The lower panel shows the average mean square error of the estimated standard errors compared to the sample standard deviations of the estimates, over all common haplotypes, under HWE, HWD(1) and HWD(2), respectively. The number of SNPs is 10.

The running time was, on average, 0.04–2 s for each data set in our simulation studies. All simulations were run on a three-machine MOSIX cluster, each with a dual Pentium III 800 MHZ with 2GB RAM, running under Linux 2.2.18-mosix.

4. An example and discussion

To illustrate our proposed method, we used an actual SNP-genotype data set introduced in Section 1. The data include 18 SNPs found on gene ARHGDIB and genotyped for 44 unrelated individuals (Reich *et al.*, 2001). Among these, only 8 SNPs had complete genotype data on all 44 individuals. The remaining 10 SNPs had missing genotypes on one or more individuals. We first estimated haplotype frequencies for the entire data set. We found 20 haplotypes (results not shown). We then estimated haplotype frequencies for the subset of data with the 8 SNPs with no missing genotypes. To compare our results with Excoffier and Slatkin's EM algorithm, we also analyzed the data using the Arlequin software (Schneider *et al.*,

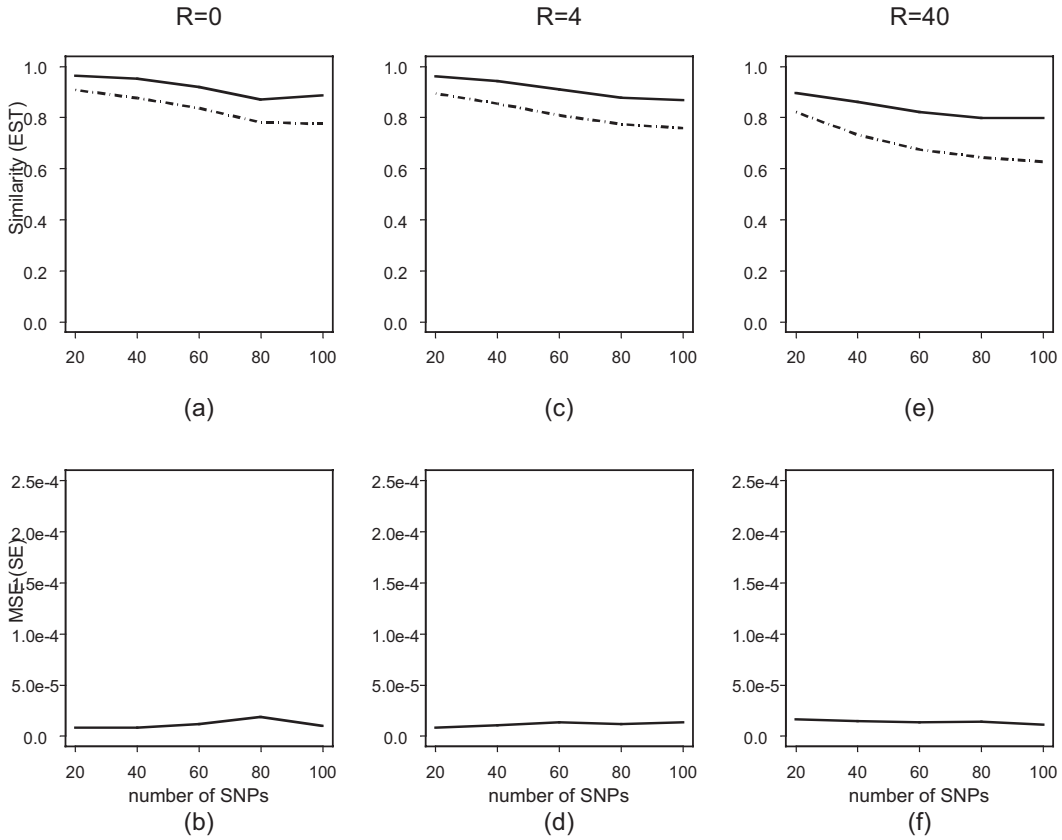

Fig. 2. The upper panel shows the similarity index I_G between the estimate and the actual haplotype frequencies in the population (dashed line) and the similarity index I_S between the estimates using our proposed method and the actual haplotype frequencies in the observed sample (solid line), for $R = 0$, 4, and 40, respectively. The lower panel shows the average mean square error of the estimated standard errors compared to the sample standard deviations of the estimates, over all common haplotypes. The sample size is 100.

2000). The estimation results are shown in Table 2. The estimation results show that haplotype frequencies obtained from both methods are identical since both are based on the same equations. The estimates of SEs from (2.5) are similar to the estimates obtained from Arlequin based on 1000 bootstraps. Since the estimation of haplotype frequencies is not part of reporting results in PHASE (implementation for Stephen *et al.*'s Bayesian method) and HAPLOTYPER (implementation for Niu *et al.*'s Bayesian method), we cannot comment on their performance, but the results should be close to the results reported in Table 2 since all four methods try to maximize the same likelihood function.

To compare the computational performance of our method (implemented in HPlus) with the Arlequin, PHASE and HAPLOTYPER implementations, we first analyzed the same subset of 8 SNPs with complete genotype data using all four implementations on the same computer. HPlus analyzed the data in under 0.08 CPU s. Compared to this, PHASE had the slowest running time at 11 min and 31 s; Arlequin took 57 s while HAPLOTYPER ran in 0.48 s—only six times less computationally efficient than HPlus. But HPlus has much larger capacity than the other implementations. To demonstrate this, our implementation analyzed genotype data sets with 632 subjects and 161 SNPs in one gene and 296 SNPs in another

Table 2. *Estimated haplotype frequencies and their standard errors of the genotype data given in Table 1 (including 8 SNPs with complete data: core SNP [G/C], G4923a6 [T/C], G4923a7 [C/G], G4923a12 [A/T], G4923a26 [A/G], G4923a44 [T/C], G4923a46 [C/T] and G4923a35 [C/T]) (allele 0: reference allele, 1: variant allele)*

Haplotype	Haplotype frequency	Standard error	
		Proposed method	EM method (with 1000 bootstraps)
10001000	0.45659	0.05113	0.05336
01110111	0.25185	0.04500	0.05020
00001000	0.12232	0.03391	0.03707
11110111	0.06570	0.02722	0.02827
01110110	0.02302	0.01591	0.01630
01110000	0.02292	0.01581	0.01607
00001111	0.02292	0.01583	0.01653
10001001	0.01166	0.01153	0.01078
01111000	0.01151	0.01138	0.01111
10000111	0.01151	0.01137	0.01098

(from the Genetic Analysis Workshop (GAW) 12), in 14 s and 2 min, respectively. We were unable to analyze these data sets with any other implementations. HAPLOTYPER, which is closest to our method in computational efficiency, is less restrictive and can handle up to 256 SNPs although it limits the number of subjects to 100.

Consequently, to compare the performance of HPlus with HAPLOTYPER, we used the GAW 12 data set with 161 SNPs and selected only the first 100 subjects. HAPLOTYPER analyzed this data set in 22 s compared to 1.3 s for HPlus. While HPlus appears to perform considerably more efficiently than HAPLOTYPER in this example, saving 20.7 s may not be important to some. However, its greater strength lies in its capacity to handle the large numbers of SNPs and/or subjects, expected in future population research.

The software was written in C++ with a user-interface and may be downloaded from our website `http://qge.fhcrc.org/hplus`.

Acknowledgements

The authors would like to thank Dr Reich and his colleagues at MIT who provided us with the SNP data used in the illustration and the associate editor and referees for their helpful comments and suggestions for revision. This work is supported, in part, by grants from NIH.

References

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N. and Nemesh, J. *et al.* (1999). Charaterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **22**, 231–238.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics* **29**, 229–232.

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

FALLIN, D. AND SCHORK, N. J. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics* **67**, 947–959.

GOLDSTEIN, D. B. (2001). Islands of linkage disequilibrium. *Nature Genetics* **29**, 109–111.

GREEN, E. D., COX, D. R. AND MYERS, R. M. (1998). The Human Genome Project and its impact on the study of human disease. In Vogelstein, B. and Kinzler, K. (eds), *The Genetic Basis of Human Cancer*, New York: McGraw-Hill, Health Professional Division 33-63.

HALUSHKA, M. K., FAN, J., BENTLEY, K., HSIE, L., SHEN, N., WEDER, A., COOPER, R., LIPSHUTZ, R. AND CHAKRAVARTI, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics* **22**, 239–247.

HAWLEY, M. E. AND KIDD, K. K. (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal Heredity* **86**, 409–411.

HUDSON, R. R. (1991). Gene genealogies and the coalescent process. In Futuyma, D. and Antonovics, J. (eds), *Oxford Surveys in Evolutionary Biology*, Vol. 7. Oxford: Oxford University Press, pp. 1–44.

HUDSON, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

KRUGLYAK, L. AND NICKERSON, D. A. (2001). Variation is the spice of life. *Nature Genetics* **27**, 234–236.

LIANG, K. Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LONG, J. C., WILLIAMS, R. C. AND URBANEK, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics* **56**, 799–810.

NIU, T., QIN, Z., XU, X. AND LIU, J. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics* **70**, 157–169.

PATIL, N., BERNO, A. J., HINDS, D, A., BARRETT, W. A., DOSHI, J. M., HACKER, C. R., KAUTZER, C. R., LEE, D. H., MARJORIBANKS, C. AND MCDONOUGH, D. P. *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723.

REICH, D. E., CARGILL, M., BOLK, S., IRELAND, J., SABETI, P. C., RICHTER, D. J., LAVERY, T., KOUYOUMJIAN, R., FARHADIAN, S. F. AND WARD, R. *et al.* (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

SCHNEIDER, S., ROESSLI, D. AND EXCOFFIER, L. (2000). Arlequin version 2.000: *A Software for Population Genetics Data Analysis*, Genetics and Biometry Laboratory, University of Geneva, Switzerland.

STEPHENS, M., SMITH, N. J. AND DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.

WANG, D. G., FAN, J., SIAO, C., BERNO, A., YOUNG, P., SAPOLSKY, R., GHANDOUR, G., PERKINS, N., WINCHESTER, E. AND SPENCER, G. *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in human genome. *Science* **280**, 1077–1082.

WIJSMAN, E. M. (1987). A deductive method of haplotype analysis in pedigrees. *American Journal of Human Genetics* **41**, 356–373.

ZEGER, S. L. AND LIANG, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

ZHAO, L. P., ARAGAKI, C., HSU, L. AND QUIAOIT, F. (1998). Mapping complex traits with single nucleotide polymorphisms. *American Journal of Human Genetics* **63**, 225–240.

[*Received April 30, 2002; first revision October 15, 2002; second revision December 18, 2002; accepted for publication December 20, 2002*]