Am. J. Hum. Genet. 67:947–959, 2000

Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data

Daniele Fallin[1] and Nicholas J. Schork[1,2,3,*]

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland; [2]Department of Biostatistics and Program for Population Genetics, Harvard University School of Public Health, Boston; and [3]The Jackson Laboratory, Bar Harbor, ME

Haplotype analyses have become increasingly common in genetic studies of human disease because of their ability to identify unique chromosomal segments likely to harbor disease-predisposing genes. The study of haplotypes is also used to investigate many population processes, such as migration and immigration rates, linkage-disequilibrium strength, and the relatedness of populations. Unfortunately, many haplotype-analysis methods require phase information that can be difficult to obtain from samples of nonhaploid species. There are, however, strategies for estimating haplotype frequencies from unphased diploid genotype data collected on a sample of individuals that make use of the expectation-maximization (EM) algorithm to overcome the missing phase information. The accuracy of such strategies, compared with other phase-determination methods, must be assessed before their use can be advocated. In this study, we consider and explore sources of error between EM-derived haplotype frequency estimates and their population parameters, noting that much of this error is due to sampling error, which is inherent in all studies, even when phase can be determined. In light of this, we focus on the additional error between haplotype frequencies within a sample data set and EM-derived haplotype frequency estimates incurred by the estimation procedure. We assess the accuracy of haplotype frequency estimation as a function of a number of factors, including sample size, number of loci studied, allele frequencies, and locus-specific allelic departures from Hardy-Weinberg and linkage equilibrium. We point out the relative impacts of sampling error and estimation error, calling attention to the pronounced accuracy of EM estimates once sampling error has been accounted for. We also suggest that many factors that may influence accuracy can be assessed empirically within a data set—a fact that can be used to create "diagnostics" that a user can turn to for assessing potential inaccuracies in estimation.

Introduction

Haplotype analyses have become increasingly popular tools for linkage-disequilibrium assessment, disease-gene discovery, genetic demography, and chromosomal-evolution studies. However, many haplotype-analysis methods rely on phase information from the individuals under study. Phase can be established by genotyping family members of each study subject to infer parental chromosomes, but this requires recruitment and genotyping of relatives, who, for many late-onset disorders, may simply not be available. An alternative involves the collection of genealogical information on all subjects to infer ancestral haplotypes, but this is again laborious. Laboratory techniques such as long-range PCR or chromosomal isolation have also been used to determine haplotypes in diploid individuals (Michalatos-Beloin et al. 1996), but these approaches are technologically demanding and often cost-prohibitive. For these reasons, haplotype-based methods have not been widely used on samples of unrelated diploid individuals, such as those typically collected as part of traditional human case-control, genetic epidemiologic, and population genetic studies.

As a solution to this problem, several rule-based and likelihood-based methods for estimating haplotype frequencies from a sample of genotyped but unphased diploid individuals have been explored, including a sequential haplotype inference algorithm (Clark 1990), and several expectation-maximization (EM)–based algorithms (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995). EM-based haplotype frequency estimates can accommodate several loci with an arbitrary number of alleles. However, analysis of a large number of loci and alleles can result in a heavy computational burden. Furthermore, most reports on the use of EM methods have not provided information on the validity of the estimates or the influence, on estimation accuracy, of population genetic factors, such as departures from

Received April 12, 2000; accepted for publication August 2, 2000; electronically published August 22, 2000.

Address for correspondence and reprints: Dr. Daniele Fallin, Department of Epidemiology and Biostatistics, Case Western Reserve University, MetroHealth Medical Campus, 2500 MetroHealth Drive, Rammelkamp Building, Room R207, Cleveland, OH 44109. E-mail: dfallin@hal.cwru.edu or (for N.J.S.) njs2@po.cwru.edu.

* Currently on leave sponsored by the GENSET Corporation of La Jolla, CA.

© 2000 by The American Society of Human Genetics. All rights reserved.
0002-9297/2000/6704-0017$02.00

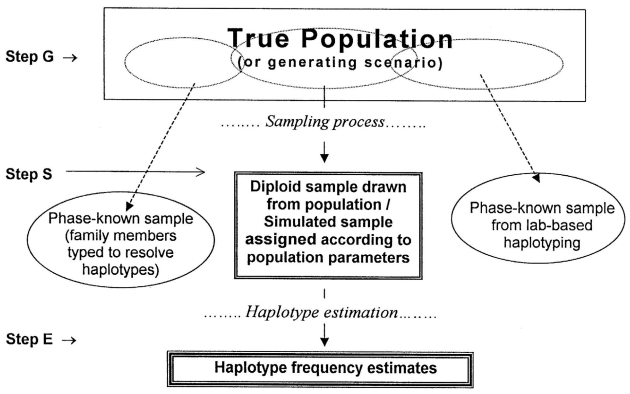

Figure 1 Conceptual framework for simulation studies and accuracy comparisons.

Hardy-Weinberg equilibrium (HWE) and actual haplotype frequency.

With the availability of single-nucleotide polymorphism (SNP) information across many genomic regions and the projected availability of dense SNP maps, haplotyping methods using this type of information will become increasingly important. With this in mind, we have developed an EM-based haplotype frequency estimation procedure tailored for biallelic data and have implemented it in a computer program that then tests for differences in haplotype frequencies between groups of individuals (N. Schork, D. Fallin, A. Cohen, L. Essioux, I. Chumakov, M. Blumenfeld, D. Cohen, unpublished data). However, before such statistics can be advocated, it is important to evaluate the accuracy of the EM-derived haplotype frequency estimates.

In this study, we consider the accuracy of the estimation procedure by measuring, through simulation, the error between EM-based haplotype frequency estimates and their true frequencies. We highlight the strong role of sampling error relative to any additional error incurred via the EM-estimation process. We then consider accuracy as a function of several population and data-set characteristics and explore the utility of data-based diagnostics for assessing probable accuracy.

Material and Methods

Our investigation of the accuracy of EM-based haplotype frequencies from unphased diploid genotype data involved simulating sample diploid data sets under different generating (or "true") population scenarios. From these generating haplotype frequency values, we drew a random sample of a specified size and then masked the haplotype resolution for each individual, simply by recording the multilocus genotypes separately, and then estimated the haplotype frequencies via our EM algorithm. This results in three main steps of the simulation-and-estimation procedure, as is shown in figure 1. We then assessed the "accuracy" of the EM frequency estimates by comparing the final estimated haplotype frequencies (E_k) to either the original generating population frequencies (G_k) or the haplotype frequencies in the sampled sets (S_k).

It is important to note the distinction between these comparisons, because they affect different issues in accuracy assessment. If the main interest is assessment the overall validity of final haplotype frequency estimates with respect to the true population values, the comparison of interest would involve estimated versus generating values (E_k vs. G_k). However, this comparison includes the effect of sampling error, which would exist even for haplotype-based methods that involve known phase, as is suggested in figure 1. A more relevant comparison for practical purposes, then, would be the accuracy of haplotype frequency estimation in relation to the haplotype frequencies from a sampled set (E_k vs. S_k), because this would only reflect any additional error caused by the estimation procedure itself. In this paper, we present both comparisons, highlight the different interpretations, and

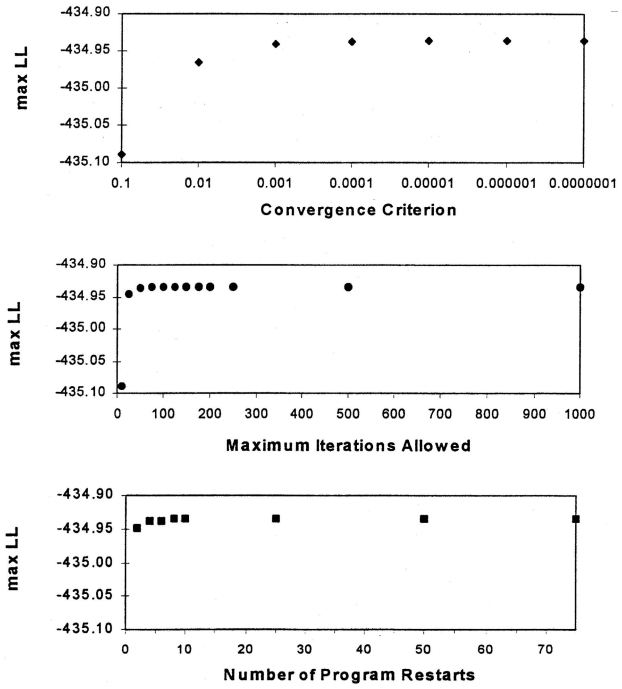

Figure 2 Distribution of maximum log-likelihoods from the estimation procedure, by program settings: convergence criterion, maximum iterations, and number of restarts at different random initial-frequency values. For these analyses, 500 data sets of 200 individuals each were simulated for a five-locus system (mean frequency .03125; variance 10.0). The analyses for each panel were performed on the same batch of 500 simulated sets each time, with the parameter of interest progressively adjusted to a more stringent value (the standard error of the maximum log-likelihood values for all situations was .098).

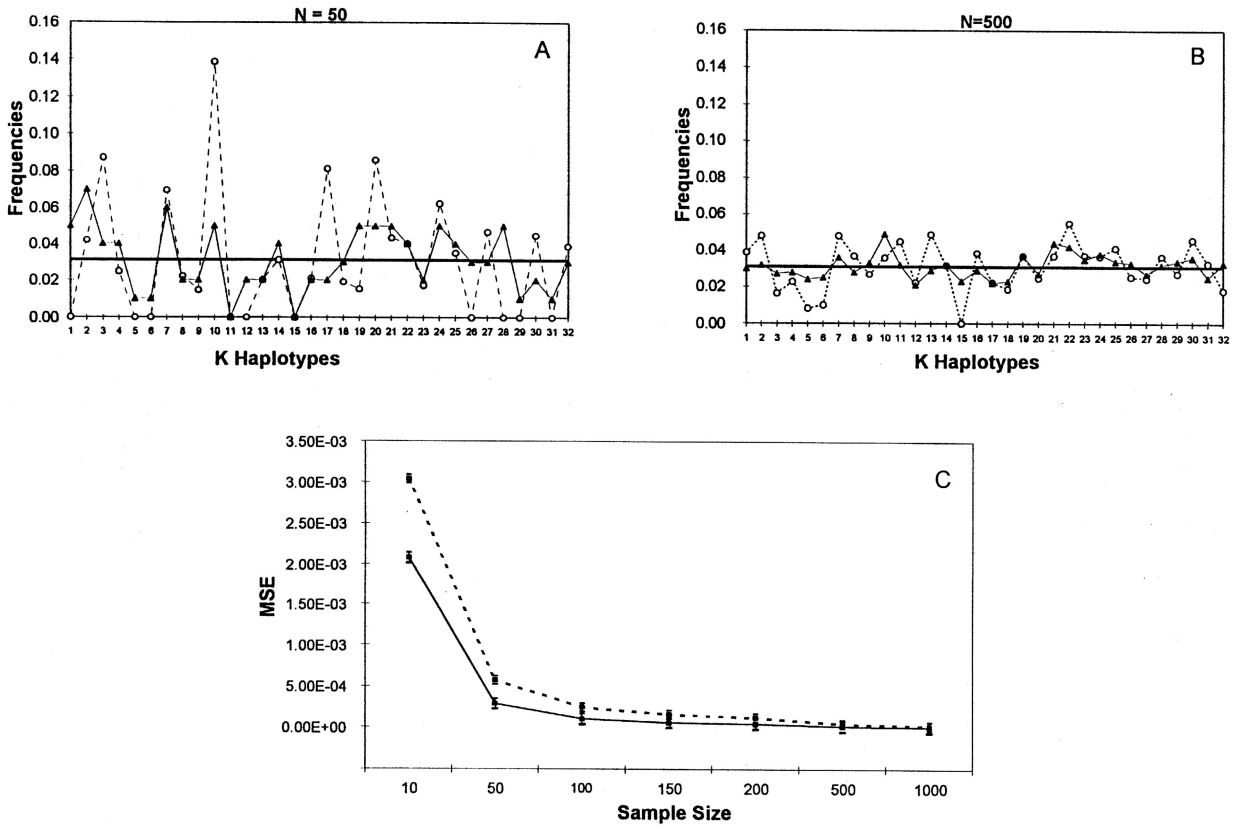

Figure 3 Influence of sample size on haplotype frequency estimates. *A* and *B*, Haplotype frequencies at the three steps of the simulation procedure. Generating frequencies (G_k [*line*]), sample frequencies (S_k [*triangles*]), and resulting haplotype frequency estimates from the EM algorithm (E_k [unblackened circles]) for a five-locus system with equally frequent population haplotype frequencies, with sample size set to $N = 50$ (*A*) and $N = 500$ (*B*) are shown. *C*, Average MSE and 95% CI for batches of 500 data sets of each sample size for five-locus haplotypes generated under the $N(1/k,\sigma^2)$ model. Unbroken line denotes comparisons of EM estimates to sample values (SE); dotted line, EM estimates to generating parameters (GE).

emphasize the roles of sampling error and possible sampling bias. We show that the additional error incurred solely by the estimation procedure is very low. In fact, the absolute difference between final frequency estimates and their true frequencies in a sample set is <.04 in most situations.

Simulated Data Sets

The first step in the simulation process involved the designation of population parameters, or generating haplotype frequencies (fig. 1, *Step G*). These generating haplotype frequencies for each data set, G_k ($k = 1,\ldots, K$; where $K = 2^L$, the number of possible haplotypes given L loci), were drawn randomly from a normal distribution with mean $1/K$ and variance $\sigma^2[N \sim (1/K,\sigma^2)]$. The generating frequencies could be constrained to be equally frequent, by setting $\sigma^2 = 0$ (i.e., $G_1 = \ldots = G_K = 1/K$), or they could be allowed to vary across all possible values between 0 and 1 by increasing the σ^2 value. We accomplished this by drawing each G_k from $N(1, \sigma^2)$ and then scaling them to sum to 1.0, by dividing the squared value of each deviate by the sum of the squared values of all deviates.

The choice of the normal distribution provides incremental departures from uniform haplotype frequencies while allowing all possible haplotype frequencies to be covered by setting the variance to be very large. This approach covered a wide variety of haplotype frequency distributions across the simulated data sets, and allowed us to measure accuracy as a function of the departure of the haplotype frequency values from uniformity (σ^2). However, to more thoroughly address the influence of haplotype frequencies on EM estimation and to ensure that haplotype frequency extremes were amply sampled, we also performed the simulations by drawing the generating haplotype frequencies from a Dirichlet distribution with a_k parameters ($k = 1,\ldots,K$). When this was employed, we used two approaches for initial a_k parameter values. First, we set each of the a_k parameters equal to 1 (uniform frequencies for each haplotype), and, second,

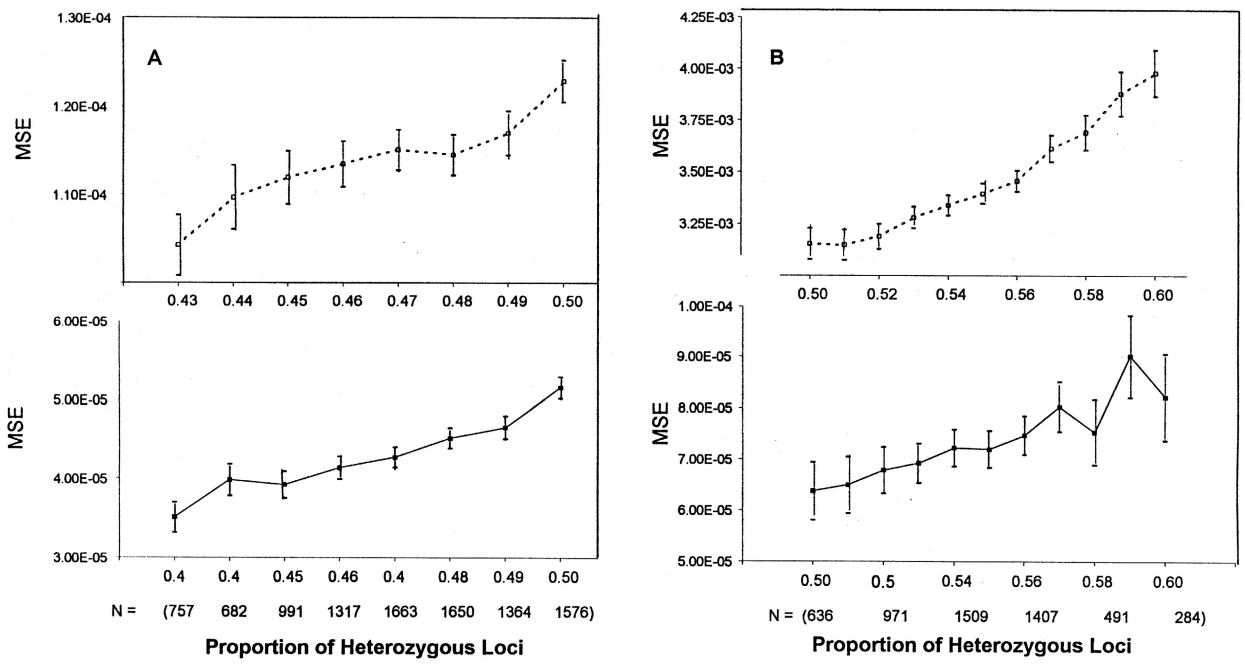

Figure 4 MSE of the final estimates as a function of the amount of missing phase information per data set. The *X*-axis indicates the proportion of heterozygous loci in the entire data set as a measure of the overall missing phase information in the sample. MSEs for the SE (*unbroken line*) and GE (*dotted line*) comparisons are plotted along the *Y*-axis. *A*, Data sets with generating haplotypes drawn for the normal distribution scenario. *B*, Data sets with generating haplotype frequencies drawn from a Dirichlet distribution with one haplotype parameter set at 50 and the rest set equal and with Hardy-Weinberg disequilibrium among the haplotypes set to .05. Both panels are based on 10,000 simulated sets (size 200 individuals), for a five-locus system with 15 restarts, 150 maximum iterations, and convergence set to 10^{-5}.

we set one parameter to be very large, relative to the rest, to ensure a dramatic dispersion of haplotype frequency values. We then measured haplotype frequency dispersion within each data set by performing a χ^2 test of the uniformity or homogeneity of resulting haplotype frequency values per simulation (denoted as a χ^2 test of uniformity).

Once the parameter values (i.e., generating haplotype frequencies) were determined by use of the methods above, a sampling procedure was done in a population with the specified generating haplotype frequency values for simulation. *N* simulated individuals were sampled (fig. 1, *Step S*) by random assignment of two haplotypes to each individual according to the generating probabilities of the *K* haplotypes.

The above sampling scenario resulted in some data sets with significant departures from Hardy-Weinberg proportions at the constituent loci. However, to ensure that ample numbers of such data sets were created to study the effect of Hardy-Weinberg disequilibrium (HWD) on estimation accuracy, we also induced HWD for some simulation batches. We assigned a first haplotype to each individual, with probabilities equal to the generating frequencies. Then the second haplotype for each individual was assigned according to the conditional probability for each of the *K* haplotypes, given the first haplotype:

$$P(H2|H1) = \frac{P(H2,H1)}{P(H1)} \quad .$$

These joint probabilities can be expressed as functions of HWD values. With this method, HWD could be induced in a way that could create more homozygosity or heterozygosity, by adjusting the strength and sign of the disequilibrium values.

The sample frequencies of each of the *K* haplotypes for each simulated data set (S_k [$k = 1,...,K$]) were then calculated by counting the number of occurrences of each haplotype in the sample and dividing by $2N$, the total number of haplotypes in a sample of *N* diploid individuals. The phase information in the resulting sample set was masked by storing the genotypes for each locus separately. These multilocus genotype data were then run through our program to produce the final estimated haplotype frequency values (E_k, $k = 1,...,K$) for accuracy comparisons (fig. 1, *Step E*).

Ultimately, our simulations were not based on a particular population-genetics model but, rather, encompassed a wide variety of allele and haplotype frequencies, allelic association strengths, and deviations from HWE scenarios. We believed that this would allow us to assess

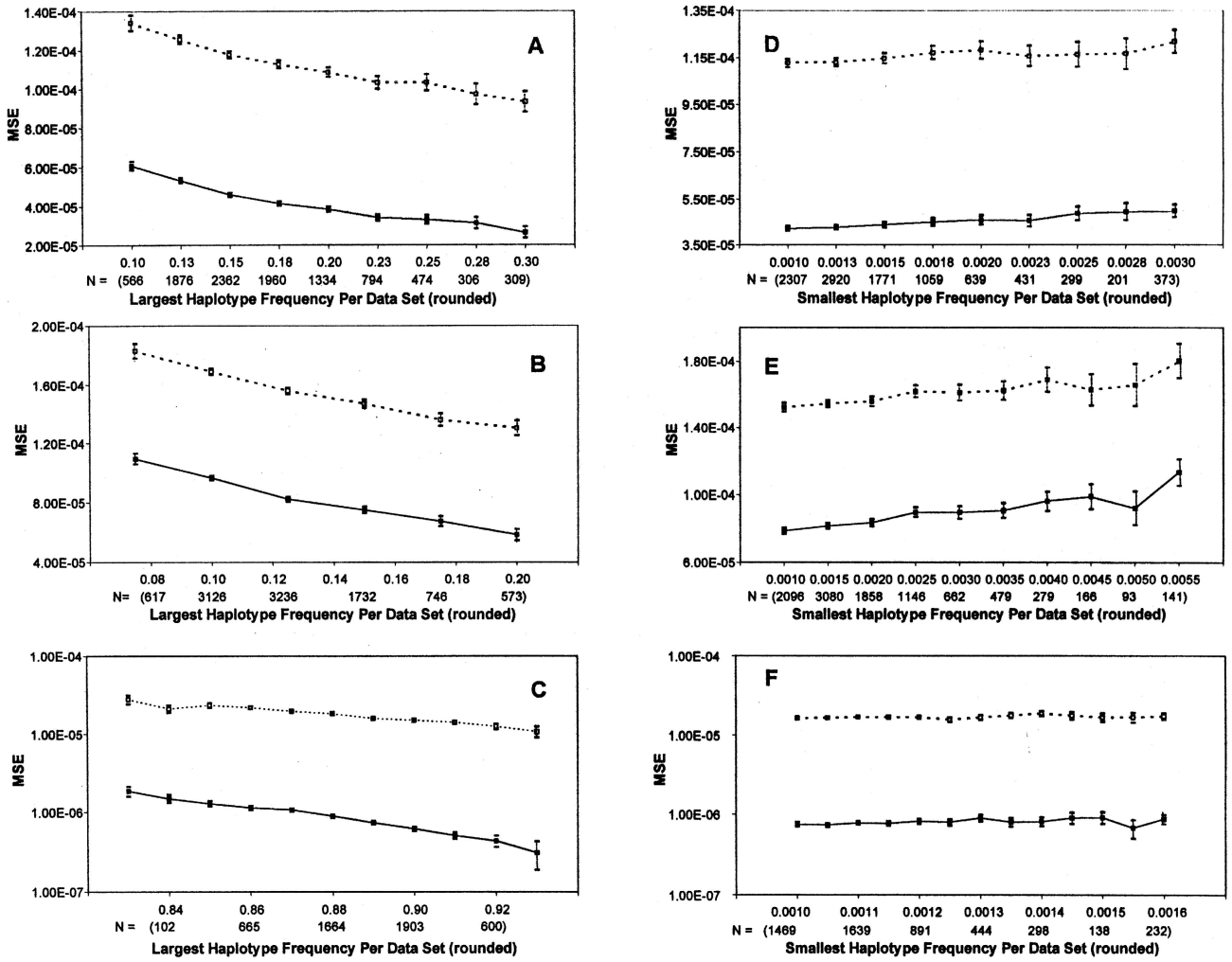

Figure 5 Accuracy of program estimates by haplotype frequency distributions within data sets. MSE measures for the SE (*unbroken line*) and GE (*dotted line*) comparisons are plotted along the *Y*-axis. In panels *A–C,* the *X*-axis indicates the frequency of the most common estimated haplotype per data set. In panels *D–F,* the *X*-axis indicates the frequency of the least common (nonzero) estimated haplotype per data set. *A* and *B,* Batches of data sets simulated under the normal generating distribution scenario. *C* and *D,* Generating haplotype frequency parameter values drawn from a Dirichlet distribution with equal parameters. *E* and *F,* Generating haplotype frequency parameter values drawn from a Dirichlet distribution with one extreme parameter value (~90%). Each panel is based on 10,000 simulated sets (size 200 individuals), for a five-locus system with 15 restarts, 150 maximum iterations, and convergence set to 10^{-5}.

accuracy in as wide a range of scenarios as possible (even some that would be considered rare and possibly unrealistic). Particular situations of interest then could be drawn from the batches of simulated data sets produced for comparison purposes.

SNP Haplotype-Estimation Program

We implemented the haplotype frequency estimation via the EM algorithm following the procedure outlined by Excoffier and Slatkin (1995). For brevity, we refer the reader to their article for details. However, the main difference between our implementation of the EM algorithm and those previously reported is that our specification is for biallelic loci only. In addition, our implementation enumerates all possible haplotypes, $H_1,...,H_K$, and diplotype configurations (i.e., specific haplotype pairs, $H_j|H_l$) consistent with each individual's multilocus genotype data and stores them throughout the iterations of the algorithm. Such storage and retrieval would be prohibitive for a large number of multiallelic loci, but they allow us to avoid having to rederive consistent diplotypes at each iteration, thereby increasing the speed of the algorithm. Another feature of our implementation of the EM algorithm is that we allow the program to be rerun automatically with different initial values to avoid convergence to local maxima. The number of such "restarts" can be

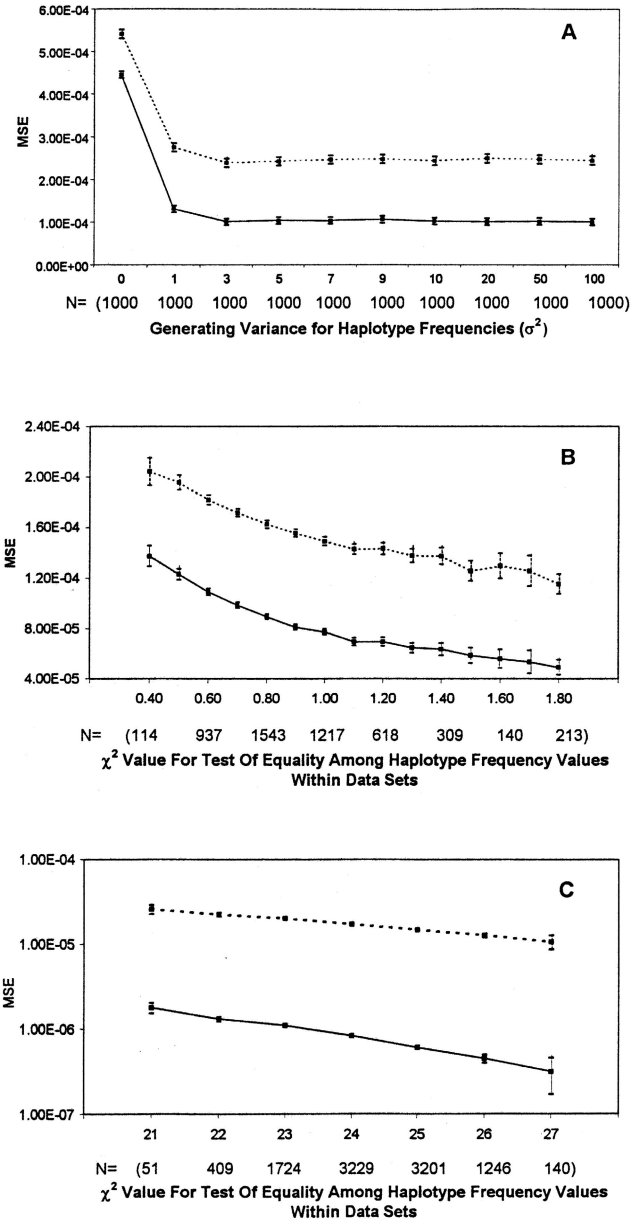

Figure 6 Accuracy of program estimates as a function of the dispersion of haplotype frequency values within a data set. MSE measures for the SE (*solid line*) and GE (*dotted line*) comparisons are plotted along the *Y*-axis. In panel *A*, the *X*-axis represents the variance used to derive generating haplotype frequency values under the normal distribution scenario. A total of 500 data sets were simulated for each variance value. In panels *B* and *C*, the *X*-axis represents the χ^2 value for a test of equality of haplotype frequency values within each data set. These panels represent batches from simulations under the Dirichlet distribution with either uniform parameters (*B*) or one extreme parameter (*C*). Both panels are based on 10,000 simulated data sets. All simulations were done for samples of 200 individuals, for a five-locus system with 15 restarts, 150 maximum iterations, and convergence set to 10^{-5}.

specified by the user, as can the convergence criterion and maximum iterations allowed per run. Our program is available for academic use by contacting one of the authors via e-mail.

Measures of Estimation Accuracy

Our primary measure of accuracy between final frequency estimates and either sample or generating values was the mean squared error (MSE). The MSE measure incorporates all K haplotype frequencies and thus captures the overall difference in haplotype frequencies between estimated and true values for a particular data set. For example, the MSE between a set of final haplotype frequency estimates and their corresponding generating values would be $\text{MSE}_{ge} = [\Sigma_k(E_k - G_k)^2]/2^L$ for $k = 1...2^L$. We plot MSE_{ge}, as well as the MSE between the final estimates and their sample-set values (MSE_{se}), as a function of the several factors under investigation. We again emphasize our focus on MSE_{se} because it measures the error attributed to the estimation procedure itself, rather than the sampling error.

As an ancillary measure of error, we also calculated the absolute difference (or absolute "bias") between the estimated frequency of each haplotype (E_k) and either its frequency in the simulated sample (S_k) or its generating value (G_k). This provides a more intuitive measure of the magnitude of error and allows for the comparison of results for different haplotype frequency values individually within and across data sets, whereas the MSE measures the composite error across all haplotype frequencies in a data set. For example, the bias between the estimated frequency of haplotype 1, E_1, and its generating population frequency, G_1, would be $B^1_{ge} = |G_1 - E_1|$. As the number of loci increases, recording this value for every possible haplotype ($K = 2^L$ possibilities) and every possible comparison would be prohibitive. As a result, we chose to calculate this absolute bias for a randomly chosen haplotype for each simulation. In addition, to measure the relative amount of error for rare versus common haplotypes, we also calculated bias for the haplotypes corresponding to the largest and smallest (nonzero) haplotype frequency values at each step (B^{gmax}; B^{gmin} for generating values, B^{smax}; B^{smin} for sample values, $B^{\text{emax}}/B^{\text{emin}}$ for final estimates). Thus, for example, $B^{\text{emax}}{}_{\text{se}} = |\text{E}_{\text{emax}} - \text{S}_{\text{emax}}|$, which reflects the difference between the final estimate and sample value for the haplotype with the largest frequency among the final estimates.

We first generated 500 data sets per simulation batch, to assess certain specifications to be used throughout the accuracy assessments. We then generated 1,000 simulations per batch for each sample size and variance parameter. We also generated 10,000 simulated data sets to examine the simultaneous effect of the investigated factors under the normal distribution scenario and then generated

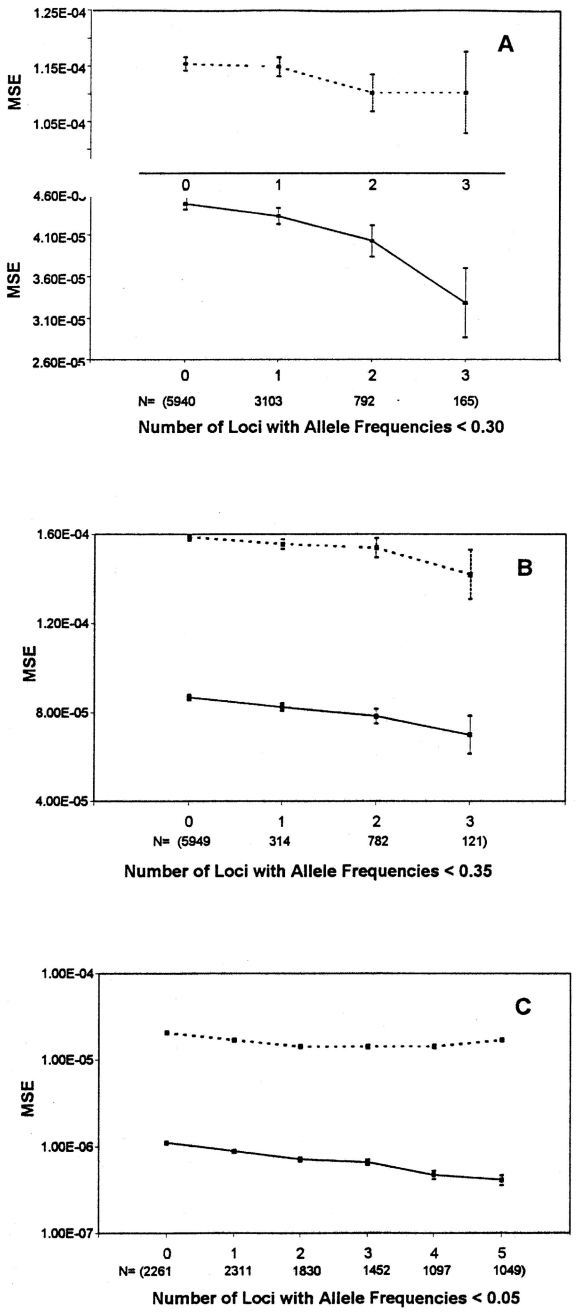

Figure 7 Accuracy of program estimates by number of constituent loci with "rare" allele frequencies per data set. MSE measures for the SE (*unbroken line*) and GE (*dotted line*) comparisons are plotted along the *Y*-axis. *A*, Batch simulated under the normal generating distribution scenario. MSE_{se} and MSE_{ge} have separate axes because of orders of scale. *B*, Generating haplotype frequency parameter values drawn from a Dirichlet distribution with equal parameters. *C*, Generating haplotype frequency parameter values drawn from a Dirichlet distribution with one extreme parameter value (~90%). Each panel is based on 10,000 simulated sets (size 200 individuals), for a five-locus system with 15 restarts, 150 maximum iterations, and convergence set to 10^{-5}.

10,000 under each of the Dirichlet scenarios described above, as well as for the HWD scenarios also discussed.

Results

Algorithm Specifications

To set optimal conditions for measuring the effect of population and data set characteristics on accuracy of haplotype frequency estimation, we first assessed the influence of several algorithm specifications. Because the EM algorithm may converge slowly and may converge to a local maximum, we examined the effect of the convergence criteria used, the maximum iterations allowed, and the number of algorithm restarts with new random initial values. Figure 2 shows the average increase and ultimate "plateau" of maximized log-likelihoods as the three parameters become increasingly stringent. From these results, setting the program to 15 restarts, the maximum number of iterations to 150, and the convergence criteria to 10^{-5} was thought to suffice for all subsequent simulations.

Sampling Error and Sample Size

We found that much of the discrepancy between the true generating or population haplotype frequencies and those estimated from the sample is due to sampling error, rather than to the estimation procedure per se. This can be seen in figures 3*A* and 3*B*, which show sample sets of sizes $N = 50$ and $N = 500$ that were drawn from equally frequent population haplotype frequencies for a five-locus system ($G_k = 1/32$ for $k = 1,...,32$). It is apparent that a large portion of the overall error (figs. 3*A* and 3*B*, circles relative to the unbroken horizontal line) is due to the sampling error between the generating and sample-set values (figs. 3*A* and 3*B*, triangles relative to unbroken line). The effect of sample size across several values is shown in figure 3*C*, in which the MSE is shown to decrease with increasing sample size. This was also reflected in the results of a regression analysis of MSE_{ge} on MSE_{gs} and MSE_{se}, in which a large fraction of the variability (~80%) in MSE_{ge} over the simulations was attributed to MSE_{gs} rather than to MSE_{se} (data not shown).

Ambiguity and Missing Phase Information

The number of "unphased" or ambiguous individuals (with respect to haplotypes) in a data set should greatly influence the accuracy of the frequency estimates, since it indicates the amount of missing phase information to be dealt with via the algorithm. Because all individuals with more than one heterozygous genotype among the loci studied are ambiguous with regard to phase, the amount of heterozygosity in a data set can be used as a proxy for the amount of missing data. Figure 4 shows the MSE measures as a function of the proportion of heterozygous

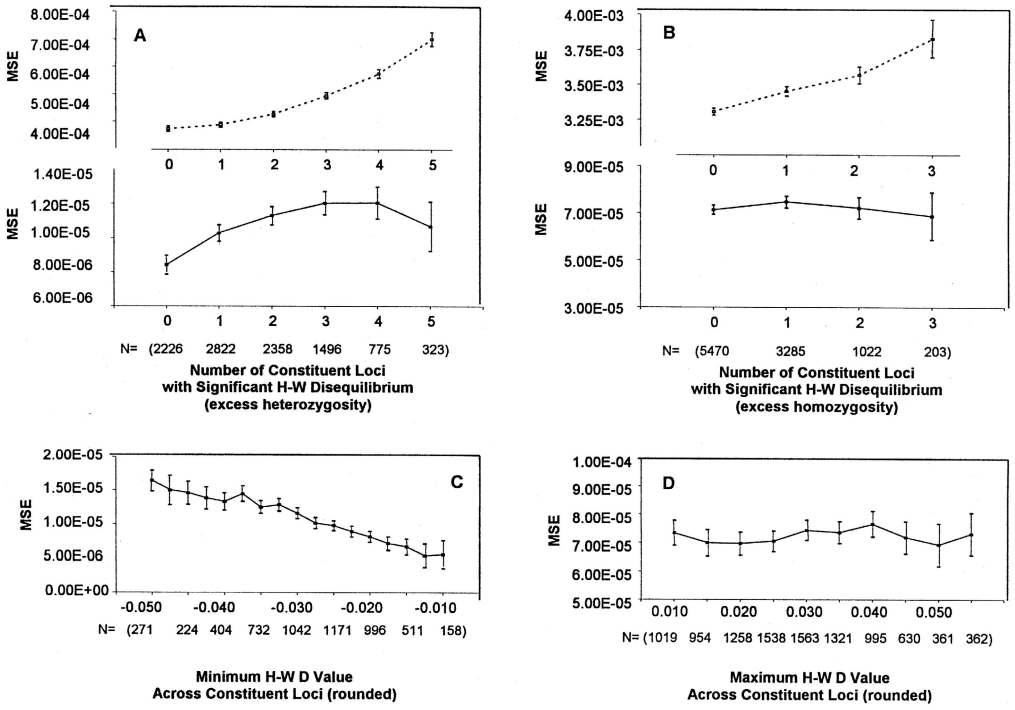

Figure 8 Accuracy of program estimates of HWD. The *Y*-axis indicates MSE between final haplotype frequency estimates and sample-set values (*unbroken line*) or generating population parameter values (*dotted line*). Each panel is based on 10,000 simulated sets under the extreme Dirichlet generating frequency model with HWD (either toward heterozygosity, [*A* and *C*] or homozygosity [*B* and *D*]) introduced at the haplotype level during the sampling process. *A*and *B*, MSE by the number of loci per simulation with significant HWD. *C* and *D*, MSE_{se} as a function of the most extreme HWD coefficient across loci per simulation. All simulations were done for 200 individuals, for a five-locus system with 15 restarts, 150 maximum iterations, and convergence set to 10^{-5}.

loci in each data set. As expected, there is a substantial increase in error as the amount of ambiguity increases, although the error associated with the greatest observed heterozygosity is still very small (i.e., ~.00012). Similar results were obtained for all of the simulating distribution strategies. In addition, the error associated with estimates, compared with their sampled values (*S* vs. *E*), is consistently lower than the overall error (*G* vs. *E*).

Haplotype Frequency Distribution

The accuracy of the estimation procedure as a function of the distribution of haplotype frequencies within a particular data set can be measured in many ways. We first describe results for the MSE values as a function of the largest and smallest haplotype frequencies within each data set, as a proxy measure for the dispersion in haplotype frequency values within each data set. As haplotype frequencies become increasingly less equal or uniform, the most common and least common frequencies will become increasingly extreme. With this in mind, the plots in figure 5 show a decrease in error with increasing maximum haplotype frequencies (panels *A*–*C*) and decreasing minimum frequencies (panels *D* and *E*). This is consistent with the notion that estimates will be better when there are some very common haplotypes and many very rare (~0 frequency) haplotypes in the population. The reason for this is that true 0-frequency haplotypes can be accurately estimated as 0, since there will be little evidence for their nonzero frequency in a data set. However, when the haplotypes are more or less equally frequent, the frequency estimates are less accurate. The only situation in which this is not directly apparent is reflected in the plot of MSE by minimum allele frequency in the extreme Dirichlet case (fig. 5*E*). In these data sets, the lack of an apparent trend in error with decreasing frequency values probably is due to the decreased range of minimum haplotype frequency values under the extreme-frequencies simulations. However, the overall range of error for this batch of simulations is much less than the error for the other simulations, supporting the notion that highly disparate haplotype frequencies correspond to better overall accuracy.

This point is also addressed in figure 6, in which MSE is plotted as a function of the variance value used to generate the population haplotype frequencies for the $N(1/K,\sigma^2)$-derived simulation (fig. 6*A*), as discussed in the

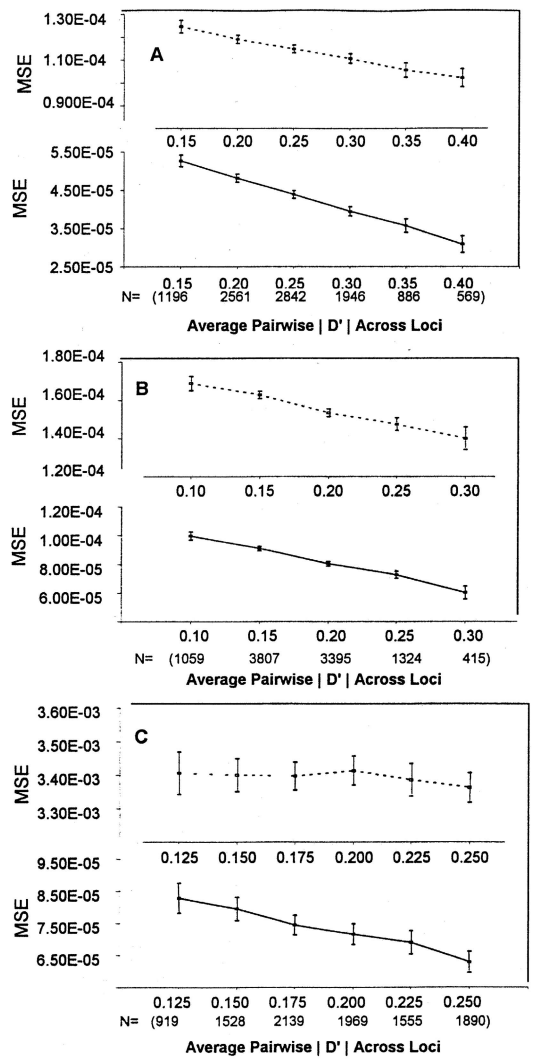

Figure 9 Accuracy of program estimates by LD. The *Y*-axis indicates MSE between final haplotype frequency estimates and sample-set values (*unbroken line*) or generating population parameter values (*dotted line*). The *X*-axis indicates the average D' LD value across all pairwise comparisons per simulation. *A*, Simulations generated under the normal distribution scenario. *B*, Population haplotype frequency values drawn from a Dirichlet distribution with equal parameters. *C*, Population haplotype frequency values drawn from a Dirichlet distribution with one extreme parameter value (~50%) and HWD induced toward homozygosity. Each panel is based on 10,000 simulated sets (size 200 individuals), for a five-locus system with 15 restarts, 150 maximum iterations, and convergence set to 10^{-5}.

Material and Methods section. For the Dirichlet-derived simulations, we performed a χ^2 test of equality of haplotype frequency values per data set such that large χ^2 values suggest greater departures from uniformity. The MSE values as a function of these χ^2 values per data set are shown in figure 6*B* and 6*C*. These plots again show a decrease in estimation error with increased dispersion (i.e., nonuniformity) of haplotype frequency values.

Haplotype Frequency Values

The above results relate to the distribution of haplotype frequencies within each data set. They do not address the amount of error to be expected for a particular haplotype frequency value. The accuracy of the EM procedure with respect to rare haplotypes is of interest to many researchers who are interested in population genetics or who are concerned with the possibility that rare haplotypes may be important for disease risk. The results above show that the EM algorithm performs well when there are very common and zero-frequency haplotypes in a population. To assess the error rates for large versus small haplotype frequency values within a particular setting, we compared the absolute bias (or absolute difference) between final estimates and their sample or generating values (B_{se} or B_{ge}) for the most frequent and least frequent (nonzero) haplotype within each data set. The bias measures for the largest and smallest haplotypes were correlated for all simulating scenarios, supporting the influence of the overall distribution of haplotype frequencies on estimation error rates. Plots and paired *t* tests of the bias values comparing estimates with true sample values for the most- and least-frequent haplotypes among the estimated values $B^{emax}{}_{se}$ and $B^{emin}{}_{se}$ suggest greater bias for the common haplotypes in all simulation scenarios (data not shown), although this was thought to be due, in part, to the relatively smaller range of the possible error values for the frequencies of rare haplotypes.

Another important measure in this regard is the number of times a rare haplotype is lost (i.e., estimated as having 0 frequency when it has a nonzero frequency in the population) because of the EM estimation. To examine this, we recorded each haplotype that was missed in the final set of estimates for each simulation. Across simulation strategies, only 0–3 haplotypes, on average, were lost between the sample sets and the EM estimates. Of these, ~90% were haplotypes with frequencies <1% (see table 1). We also recorded the number of simulations in which haplotypes of a particular value were lost among the final estimates. To monitor the number of missed haplotypes across simulations in this way, we tallied the frequency of simulations per batch in which a rare haplotype (defined as that with the smallest frequency >.001 [or .01 in some analyses]) was missed in the final estimates. As might be expected, rare haplotypes with population frequencies of ~.1% were lost in the final estimation in 64%–83% of the 10,000 simulations (see table 1). The number of simulations in which rare haplotypes among the sample data (defined with the same minimum threshold: smallest frequency $\geq$.001) were lost in the final estimates was lower in all simulation scenarios (table 1, column 3). This again shows the importance of sampling error in haplotype frequency estimation. In fact, in the extreme Dirichlet

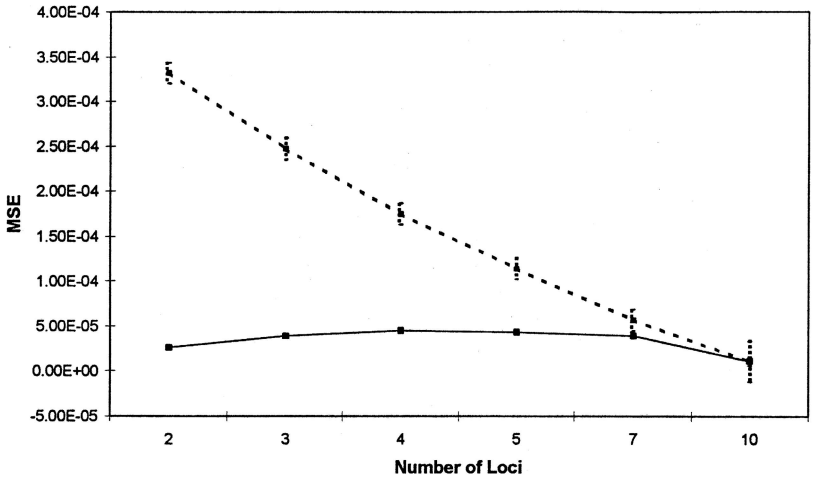

Figure 10 Accuracy of program estimates by number of loci in haplotype. The *Y*-axis indicates MSE between final haplotype frequency estimates and sample-set values (*unbroken line*) or generating population parameter values (*dotted line*). The categories represented on the *X*-axis correspond to 1,000 simulations each (2-, 3-, 4-, 5-, 7-, and 10-locus systems). All simulations were for 200 individuals, with 15 restarts, 150 maximum iterations, and convergence set to 10^{-5}.

scenario, in which one haplotype is very frequent and the rest are very rare, 67% of the simulations lost the "rare" haplotype designated from the population set, but only 8% lost the rare haplotype defined from the haplotypes available in the sample set. This suggests that the difficulty in determining rare haplotype frequencies is often a sampling problem, rather than an error in the frequency-estimation procedure: if the haplotype is never sampled, it cannot be estimated well. If the threshold of "rare" is changed to $\geq$1%, the proportion of simulations losing the rare haplotype from the population or sample sets decreases dramatically. Again, the proportion of loss of rare population haplotypes is greater than the loss of rare sample haplotypes.

Allele Frequencies

Given the results for haplotype frequency distributions within data sets, it would be expected that allele frequencies at the relevant loci may predict accuracy levels, since rare alleles at biallelic loci will create some very rare (and, thus, some very common) haplotypes. To assess this, we tallied the number of loci within each data set with a minor-allele frequency below a particular threshold (designated as "rare" for those simulations). Figure 7 shows plots of the MSE measures as a function of the number of constituent loci with "rare" alleles (defined by a lower frequency threshold particular to each simulation batch). Data sets with increased numbers of "rare" alleles showed increased accuracy in all simulating scenarios. We also calculated the minimum and the average minor-allele frequency across all constituent loci per data set and plotted these by MSE and bias. These also showed a decrease in error with decreasing minor allele frequencies (figures not shown).

Departure from HWE

Departure from HWE may be a substantial source of error in EM haplotype frequency estimation, simply because the algorithm relies on HWE in its "expectation" step. Thus, one might expect to see a loss of estimation accuracy when alleles at the loci are not in HWE. However, departures from HWE may result in an excess homozygosity, which could, in effect, decrease the amount of ambiguous phase information in a data set and, as such, improve estimation accuracy. To assess the effect of HWE departures, especially with respect to excess homozygosity versus heterozygosity, we induced positive and negative HWD (toward homozygosity or heterozygosity) among the sampled haplotypes in the simulation procedure (as described in the Material and Methods section, above). We calculated the HWD coefficient, D, (Weir 1996) and χ^2 statistics for a test of HWE at each locus for each simulation. Figure 8 plots MSE versus two measures of HWD: excess heterozygosity (panels *A* and *C*) and excess homozygosity (panels *B* and *D*). The first two panels (fig. 8*A* and 8*B*) show the effect of increasing numbers of loci per simulation, with significant departures from HWE (as measured by χ^2 values >3.84) on MSE. In panel *A*, there is a clear increase in error as the number of loci showing HWD (toward excess heterozygosity) increases, reflecting the influence of heterozygosity on estimation accuracy. The corresponding plot of error be-

Table 1

Proportion of 10,000 Simulations in Which the Designated "Rare" Haplotype Was Lost in the Final Haplotype Frequency Estimates, and the Average Number of Haplotypes Missed per Simulation

	% of Simulations in Which the "Rare" Haplotype Was Lost				Average No. of Sampled Haplotypes Missed per Simulation	% of Missed Haplotypes with Frequencies <.01
	Smallest Frequency ⩾.001		Smallest Frequency ⩾.01			
Generating Distribution for Batches	Population Minimum Haplotype	Sample Minimum Haplotype	Population Minimum Haplotype	Sample Minimum Haplotype		
Normal	68.27	50.72	15.22	9.65	3	97
Dirichlet[a]	64.46	54.99	22.35	17	3	88
Dirichlet[b]	66.89	7.45	2.11	0	0	100
Dirichlet[c]	82.91	60.79	7.11	8.67	2	93

[a] Population frequencies drawn from Dirichlet distribution with equal parameter values.
[b] Population frequencies drawn from Dirichlet distribution with one extreme parameter value.
[c] Population frequencies drawn from Dirichlet distribution with one extreme parameter value and HWD ($D = .05$) induced between haplotypes during sampling.

tween the EM estimates and their sample values in panel *B* shows no increase in error with more-extreme HWD values (towards excess homozygosity). This may indicate the relative increases in phase information as loci depart from HWE. The increase in overall error with increasing homozygosity, in comparison, may reflect sampling error. Increased departures from HWE may result in the loss of some haplotypes in the sample set and thus may increase error. This again emphasizes the role of sampling error, rather than estimation error, in accuracy. Panels *C* and *D* of figure 8 show MSE between EM estimates and their sample values as a function of the most extreme HWD coefficient across the constituent loci for each simulation. In accordance with the results for the number of loci not in HWE, there is a clear increase in error as D coefficients become more negative (i.e., toward excess heterozygosity [panel *C*]), while this effect is not seen for increasingly positive D values (toward homozygosity).

Linkage Disequilibrium

The amount of linkage disequilibrium (LD) between loci should also have an important effect on the haplotype-estimation accuracy, since haplotypes should be more uniformly distributed for loci in complete equilibrium and roughly equal allele frequencies. We assessed LD by computing Lewontin's (1964) D' for all pairs of loci as well as χ^2 statistics gauging the significance of the LD. Plots of the accuracy measures by the D' values are shown in figure 9. In the three simulation scenarios shown, the error between estimates and their sample frequencies decreases and the average LD across the constituent loci becomes stronger. The same effect is seen for the overall error in the normal and uniform Dirichlet simulating scenarios. For the extreme Dirichlet case, the overall error does not show an obvious trend, probably because of the added effect of sampling error. Plots assessing the average χ^2 value for tests of pairwise LD also showed this pattern (plots not shown).

Number of Loci

Figure 10 shows an overall increase in accuracy as the number of loci analyzed increases when accuracy is assessed relative to their generating values. However, we found that estimation error appears to peak, for four- or five-locus systems, when final frequency estimates are compared with sample values. This initial increase in estimation error may be expected, given the increase in loci and the corresponding increase in possible haplotypes to be estimated for the sample size, and could reflect an increase in missing data. The decline in sample-to-estimate error with a larger number of loci (>5) may be somewhat artificial. Consider the fact that a large number of loci likely will generate a large number of possible haplotypes, many with 0 frequency. Samples with many 0-frequency haplotypes would result in frequency estimates of the haplotypes as 0, and the MSE across them would be low as a result.

Multiple Factors

In an attempt to assess the relative importance of all the factors studied with respect to the accuracy of haplotype frequency estimation, we entered the factors into a regression model predicting the mean squared error between final estimates and their frequencies among the sample set for each simulation (MSE_{se}). Table 2 displays the results of single-factor analyses, as well as those of this multiple-regression approach, for the 10,000 simulations under the normal generating scenario. Because most of the factors we investigated can be assessed within a particular data set of interest, such quantification of the

Table 2

Regression of MSE between Estimated Frequencies and Their Sample Frequencies (MSEse) on All Factors

	Single-Factor Regression				Multifactor Regression		
Factor[a]	Correlation	R^2	F	$P>F$	Adjusted Coefficient	T	$P>\|T\|$
Constant	...	...	...	...	...	12.29504	1.96E-23
Proportion of heterozygotes	.16944	.029	295.531	<.001	.036430	2.468120	.013599
χ^2 for equality of haplotype frequencies	−.37370	.140	1622.86	<.001	−.436520	−35.78120	1.96E-23
No. of loci with "rare" alleles	−.06383	.004	40.902	<.001	.141553	10.79422	1.96E-23
Maximum HWD value	−.11192	.013	126.825	<.001	−.105800	−9.295550	1.77E-20
Average LD D' value	−.21809	.048	499.281	<.001	.021182	1.881338	.059955

[a] Analyses were performed on 10,000 simulations under the normal-distribution generating model.

relative predictive value of each factor may allow researchers to assess how likely a particular data set is to provide reliable haplotype frequency estimates. From the full-model regression results, as well as from univariate analyses, it appears that all factors show a significant trend in the direction observed in the relevant plots of figures 4–9. The dispersion of haplotypes within a data set appears to be the strongest predictor from the univariate and multiple-regression models.

Discussion

We have demonstrated, via extensive simulation studies, that haplotype frequency estimation for biallelic diploid genotype samples via the EM algorithm performs very well under a wide range of population and data-set scenarios. In fact, even the worst haplotype frequency estimates from our studies were highly accurate (for five-locus haplotypes, 60% of the estimates lie within 3% of their generating values, and 96% lie within 6% of their generating values). Ultimately, our studies suggest that much of the overall error between the original population parameters and the final frequency estimates is due to sampling error, rather than to algorithmic and estimation problems or inaccuracies. This is supported by the increase in overall accuracy with increasing sample size. This point deserves emphasis, because it implies that greater attention should be paid to the sampling scheme for all haplotype-based study designs. The additional error incurred via estimation, versus some other form of phase determination, is relatively minor in comparison. An improvement in accuracy of the estimation procedure itself (as measured by sample-estimate [ES] comparisons) with increased sample size was also observed. This is likely a function of several factors. The EM algorithm we used assumes HWE, and larger sample sizes provide better representation of HWE, if it truly exists in the source population. For this reason, both the population-sample (SG) and ES error levels decreased with large sample sizes. In addition, the algorithm works best with low amounts of "ambiguous" individuals (i.e., individuals with unresolvable phase information), and larger sample sizes also provide a greater number of unambiguous individuals, resulting in a further decrease in error between the sample and estimated values with increasing sample size.

We also find that the most influential effect on estimation accuracy is the dispersion of allele and haplotype frequency values within a data set. As the haplotype frequencies become more unequal, the more-frequent haplotypes can be estimated accurately. In addition, when many haplotypes have zero frequency, their absence in the data set will generally allow accurate estimation of this zero frequency, contributing to a small overall error in frequency estimation.

In addition to the effect of overall dispersion of haplotype frequencies within a data set, we also examined the relative accuracy of the estimation of rare alleles versus common alleles. Our results show that, although it is true that very rare haplotypes (frequency in population $\leqslant$.1%) are often lost among the final estimates (i.e., estimated as frequency 0), this may be more a result of sampling error (because these haplotypes were often not included in the sample) than of EM estimation. If the primary interest is determination of disease-predisposing haplotypes, it may be argued that haplotypes not seen in the case sample are unlikely to have contributed to disease status among cases. In this vein, the concern would be the extent to which rare disease-predisposing haplotypes among the sampled individuals are missed because of the estimation procedure. Although we show that this may be a problem for very rare haplotypes among the sampled individuals (frequency in the sample $\leqslant$.1%), the likelihood that such extremely rare haplotypes contribute appreciably to the risk of disease among the affected individuals in the sample is very low. Haplotypes with frequencies >1% among the sampled individuals do not tend to be lost that frequently, suggesting that haplotype estimation for genetic epidemiological studies may be a viable method. Ultimately, this will be a function of the particular disease predisposition in question, but, for most common disorders, we emphasize that the impor-

tance of the error incurred via phase estimation is likely to be much less than that of sampling strategy.

The influence of departures from HWE on estimation accuracy emphasizes the importance of the directionality of such disequilibrium. It might be expected that departures from HWE, given the EM algorithm's exploitation of HWE to compute expected haplotype frequencies, would significantly influence the accuracy of the resulting estimates. However, as was noted recently by Osier and colleagues (1999), there is balance between loss of accuracy caused by departure from HWE and gain of accuracy caused by the decrease in missing phase information with an excess of homozygosity, such as might result from departures from HWE. Our studies bear this out as well, since departures from HWE that result in an excess heterozygosity do lose accuracy, whereas those that result in an excess homozygosity do not. This issue is of particular relevance when one has sampled diseased individuals, since an excess homozygosity may be expected at the disease allele and all those in LD with it, especially if the disease is recessive (Nielsen et al. 1998).

Our use of a regression model to assess the simultaneous effect of different factors on estimated accuracy may have some utility. Many of the factors that we studied in terms of their influence on accuracy can be assessed within a given data set (e.g., evidence for departure from HWE, number of heterozygous genotypes, pairwise LD among the loci, relative frequency of haplotypes among the final estimates, etc.). Thus, through the regression-model outcomes, one might be able to predict MSE or bias from their own data. The results of this prediction could then serve as a "diagnostic" for potential inaccuracies in haplotype frequency estimates caused by features in the relevant data set. We intend this analysis to be an initial example of the possibility for such diagnostics within an observed data set, and we emphasize that this regression approach is not without problems. For example, we assumed linearity and did not consider interaction effects between the factors, which may be justified in our model, given the relationship between the factors. Furthermore, possible biases could be inherent in the simulated data sets, which were not explored. The single-factor regressions do provide a measure of the relative importance of each factor in predicting error caused by the estimation procedure, and they also provide tests of significance of the trends plotted in the figures presented. More work regarding the utility of the multiple-regression approach is needed, however.

Ultimately, the results of our studies suggest that even in the worst cases, individual haplotype frequency estimates via the EM algorithm do not deviate much beyond 5% of their true value among sampled individuals for sample sizes $\geqslant$100. Researchers should be concerned with the quality of sampling to a much greater extent than the possibility for estimation errors when assessing haplotypes among unphased individuals. In light of this, EM estimation of haplotype frequencies for multiple diallelic genotypes may be a viable alternative to the recruitment of additional family members or intensive laboratory haplotyping for haplotype-based genetic studies. Finally, we would like to emphasize that the results in this study refer to the accuracy of haplotype frequency estimation only. The extent to which the factors we studied influence any statistical-inference procedures that make use of haplotype frequency estimates demands independent attention. We are currently pursuing such topics for future publication.

Acknowledgments

We would like to thank the reviewers of this manuscript for their thoughtful suggestions. This work was supported in part by NIH grants HL94-011 and HL54998-01 (awarded to N.J.S.).

References

Clark A (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122

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

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

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49: 49–67

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

Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 24:4841–4843

Nielsen DM, Ehm MG, Weir BS (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 63:1531–1540

Osier M, Pakstis A, Kidd JR, Lee JF, Yin SJ, Ko HC, Edenberg H, Lu RB, Kidd KK (1999) Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. Am J Hum Genet 64:1147–1157

Schork N, Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D (2000) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res, submitted

Weir BS (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA