correctly reconstructed individuals, an individual may be reconstructed correctly by the Phase method but not by the EM method; on the other hand, an individual may be reconstructed correctly by the EM method but not by the Phase method.

In the present study, we have compared the EM method with a recently proposed haplotype reconstruction method (Stephens et al. 2001), through use of empirical population haplotype frequency data and phase-known genotype data sets. The PHASE method is based on the coalescent theory; however, it is likely that a simple coalescent model will not be a good representation of the actual history of a human population because of fluctuating population size, migration, and other factors. If the model is not appropriate, analyses that assume the model cannot be expected to yield more-accurate estimates of haplotype frequencies than analyses making no historical assumptions. The degree to which such a model is representative may vary according to population and locus. In the results of our simulations using empirical population haplotype frequency data, the PHASE method showed no improvements over the EM method, except at the RET locus in an African population. For the nine African populations in which haplotypes were inferred through molecular methods, the EM method and the PHASE method yielded almost identical results in seven populations, and the PHASE method did outperform the EM method in the other two populations. Therefore, our systematic comparisons suggest that the PHASE method may not yield consistently significantly improved estimates; this is contrary to the consistent improvements observed by Stephens et al. (2001). In summary, across all of the examples studied, the PHASE method did not yield significantly different results from a simple maximum-likelihood procedure.

Acknowledgments

This research was supported, in part, by National Institutes of Health grants GM59507 and HD36834 (to H.Z.) and GM57672 and AA09379 (to K.K.K.).

SHUANGLIN ZHANG,[1] ANDREW J. PAKSTIS,[2] KENNETH K. KIDD,[2] AND HONGYU ZHAO[1,2]
Departments of [1]Epidemiology and Public Health and [2]Genetics
Yale University School of Medicine
New Haven

Electronic-Database Information

The URL for data in this article is as follows:

ALFRED Web site, http://alfred.med.yale.edu/alfred/index.asp (for population descriptions and haplotype definitions)

References

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

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

Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK (2001) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms—an update. Nucleic Acids Res 29:317–319

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

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518–522

Xie X, Ott J (1993) Testing linkage disequilibrium between a disease gene and marker loci. Am J Hum Genet Suppl 53: 1107

Address for correspondence and reprints: Dr. Hongyu Zhao, Department of Epidemiology and Public Health, 60 College Street, Yale University School of Medicine, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

0002-9297/2001/6904-0028$02.00

Am. J. Hum. Genet. 69:912–914, 2001

Reply to Zhang et al.

To the Editor:

Stephens et al. (2001) (henceforth referred to as "SSD") introduced a new statistical method for haplotype reconstruction, called "PHASE," that has three major advantages over existing approaches, including EM. The letter from Zhang et al. (2001 [in this issue]) (henceforth referred to as "ZPKZ"), questions one of these—namely, the increased accuracy of PHASE.

ZPKZ report two kinds of comparisons. The first is based on "empirical population haplotype frequency data," and the second is based on data for which the true phase is determined experimentally. Only the second of these types is actually based on "real" data in the usual sense, and when these data are used, PHASE does considerably outperform EM. We report comparisons below, using three other real data sets. In each case, PHASE provides haplotype reconstructions that are more accurate than those provided by EM, sometimes considerably so.

Table 1

Discrepancies Obtained by PHASE and EM on Genotypes from the CAPN10 Locus

	Discrepancy Obtained by	
Sample	EM Method	PHASE Method
Combined	.13	.05
Population 1	.14	.09
Population 2	.26	.08
Population 3	.00	.00
Population 4	.23	.13

Much of the discussion by ZPKZ—as well as, apparently, their discouraging conclusion for PHASE—is based on their first set of comparisons. Unfortunately, their terminology may cause some confusion. The "empirical" haplotype frequencies on which they base their comparisons are not, in fact, haplotype counts in real data. Instead (S. Zhang, personal communication), although not mentioned in their letter, the "empirical" frequencies are actually estimates, provided by the EM algorithm, from genotype data.

PHASE is best thought of as a Bayesian method for haplotype reconstruction. Its potential to improve on maximum likelihood (and, hence, on EM) comes from its use of prior information. In particular, it incorporates the prior knowledge that unresolved haplotypes will tend to be the same as, *or similar to,* known haplotypes. When this is true in actual data, PHASE will typically provide better haplotype estimates. The comparisons by ZPKZ suggest that, when such clustering of haplotypes is not present, PHASE does not perform systematically worse than EM.

As emphasized by SSD, although PHASE uses a coalescent approximation to quantify the fact that haplotypes tend to be similar to one another, PHASE does not *depend* on the assumptions underlying the coalescent model, and we would expect it to perform well under much more general settings, including population structure, recombination, and selection.

In collaboration with H. Ackerman, we have compared EM and PHASE for haplotypes determined from pedigree data at the IL8 and TNF loci. At the IL8 locus, Hull et al. (2001) typed six single-nucleotide polymorphisms (SNPs) over 4.5 kb in 61 Gambian parents-child triples. Of the 122 parents, 102 had haplotypes that were unambiguous or that could be determined from the child's genotype. At the TNF locus, H. Ackerman (unpublished data) typed 12 SNPs over 4.3 kb in 53 Gambian parents-child triples, and the same procedure gave 96 unambiguous parents. For each locus, we applied EM and PHASE to the subset of unambiguous parents and computed the error rates. At IL8, error rates were 7/31 for EM and 6/31 for PHASE; at TNF, error rates were 24/88 for EM and 10/88 for PHASE. Thus, PHASE reduced error rates in these data sets by 14% and 58%, respectively.

We are grateful to S.M. Fullerton, G. Ybazeta, and A. DiRienzo (personal communication), for allowing us to report the following results of their unpublished comparison of PHASE and EM on molecularly determined haplotypes at the CAPN10 locus. They typed 46 individuals from four populations (n = 11, 12, 11, and 12) at 14 biallelic SNPs and found the discrepancy for the algorithms applied to the combined sample and applied to the four population samples separately. PHASE consistently outperformed EM, reducing discrepancy by as much as 69% (table 1).

In summary:

1. PHASE typically provides more-accurate haplotype estimates than does EM and other existing methods, when there is "clustering" in the true haplotype configuration.
2. Such clustering would usually be expected in real data, on population genetics grounds, whether or not the data are well modelled by the standard coalescent.
3. PHASE outperforms EM for the one real data set in ZPKZ and for the three other real data sets we have looked at.
4. Most of the comparisons by ZPKZ are based not on real haplotype data but rather on genotype data from which haplotype frequencies have been estimated by EM. Haplotype frequencies estimated by EM will not necessarily exhibit clustering, even if it is present in the true frequencies. It is thus not surprising—and, perhaps, not directly relevant—that, in most instances, ZPKZ observe similar behavior between EM and PHASE.
5. When the true haplotypes do not exhibit clustering, PHASE does not seem to perform systematically worse than EM.

Thus, although we admit that there will be exceptions, PHASE provides more-accurate haplotype reconstructions than EM for all the real data sets we and ZPKZ have examined and under conditions which seem likely for most other real data sets. In other settings, it performs no worse. In this sense, using PHASE is a low-risk strategy with considerable potential gains; however, increased accuracy is only one of the advantages of PHASE. We continue to regard the other advantages as being at least as important. It remains the case that PHASE is practicable for much larger problems than is EM, and it is the only available method that provides an accurate measure of the uncertainty associated with phase calls, thus guarding against inappropriate overconfidence in statistically reconstructed haplotypes.

MATTHEW STEPHENS,[1] NICHOLAS J. SMITH,[2] AND PETER DONNELLY[3]
[1]*Department of Statistics, University of Washington, Seattle; and Departments of* [2]*Biochemistry and* [3]*Statistics, University of Oxford, Oxford, United Kingdom*

References

Hull J, Ackerman H, Isles K, Usen S, Pinder M, Thomson A, Kwiatkowski D (2001) Unusual haplotypic structure of *IL8*, a susceptibility locus for a common respiratory virus. Am J Hum Genet 69:413–419

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

Zhang S, Pakstis AJ, Kidd KK, Zhao H (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 69:906–912 (in this issue)

Address for correspondence and reprints: Dr. Matthew Stephens, Department of Statistics, University of Washington, Box #354322, Seattle, WA 98195-4322. E-mail: stephens@stat.washington.edu

0002-9297/2001/6904-0029$02.00