References

Luiselli D, Simoni L, Tarazona-Santos E, Pastor S, Pettener D (2000) Genetic structure of Quechua-speakers of Central Andes and geographic patterns of gene frequencies in South Amerindian populations. Am J Phys Anthropol 113:5–17

Rothhammer F, Moraga M (2001) Patterns of Y-chromosome variation in South Amerindians. Am J Hum Genet 69:904 (in this issue)

Rothhammer F, Silva C (1989) Peopling of Andean South America. Am J Phys Anthropol 78:403–410

——— (1992) Gene geography of South America: testing models of population displacement based on archeological evidence. Am J Phys Anthropol 89:441–446

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000*a*) Reconstruction of prehistory on the basis of genetic data. Am J Hum Genet 66:1177–1179

Simoni L, Tarazona-Santos E, Luiselli D, Pettener D (2000*b*) Genetic differentiation of South America native populations inferred from classical markers: from explorative analysis to a working hypothesis. In: Renfrew C (ed) America past, America present: genes and languages in the Americas and beyond. McDonald Institute for Archeological Research, Cambridge, pp 123–134

Sokal RR, Oden NL, Thomson BA (1999) A problem with synthetic maps. Hum Biol 71:1–13

Tarazona-Santos E, Carvalho-Silva D, Pettener D, Luiselli D, De Stefano GF, Martinez-Labarga C, Rickards O, Tyler-Smith C, Pena SDJ, Santos FR (2001) Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am J Hum Genet 68:1485–1496

Address for correspondence and reprints: Dr. Fabrício R. Santos. Departamento de Biologia Geral, ICB, UFMG, Av. Antônio Carlos 6627, CP 486, 31.270-010, Belo Horizonte, MG, Brazil. E-mail: fsantos@ icb.ufmg.br

* Present affiliation: Department of Biology, University of Maryland, College Park, MD.

0002-9297/2001/6904-0027$02.00

Am. J. Hum. Genet. 69:906–912, 2001

Comparisons of Two Methods for Haplotype Reconstruction and Haplotype Frequency Estimation from Population Data

To the Editor:

Haplotype reconstruction is an important issue, both in population genetics and in the identification of complex disease genes. Stephens et al. (2001) proposed a new statistical method (called the "PHASE method" in the following discussion, after the name of their computer program) for haplotype reconstruction based on phase-unknown marker genotype data from unrelated individuals in a population. On the basis of their simulations using coalescent models, they found that the PHASE method can reduce the error rate by >50% relative to the maximum-likelihood method, implemented via the expectation-maximization (EM) algorithm (Xie and Ott 1993; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995). One limitation of their study is the fact that their simulations are based on coalescent models, which may not be good approximations of human population evolutionary histories. In fact, the authors acknowledge that "there simply do not exist enough real data sets, with known haplotypes for sequence or closely linked markers, to allow sensible statistical comparisons of different methods" (Stephens et al. 2001; p. 982). In this letter, we report a comparison of the two methods; our comparisons involve phase-known genotype data sets, as well as simulations using empirical population haplotype frequency data. Our results show that, in general, for most of the populations studied, there is no significant difference between the PHASE method and the EM method, both in the average error rate for haplotype reconstruction and in the discrepancy (see the report by Stephens et al. [2001] for definitions of these measures) between the estimated and true sample haplotype frequencies.

For our simulations based on empirical population haplotype frequency data, we used population haplotype frequencies for four loci (RET, COMT, HOXB and D4S10, with 3, 4, 5, and 6 polymorphisms, respectively) found in samples of four populations: European Americans, San Francisco Chinese, Biaka, and Maya. We use these four populations to represent the populations from four different continents. Descriptions of the populations and of the samples of those populations, as well as the haplotype definitions, can be found in ALFRED (Osier et al. 2001; ALFRED Web site). For each locus and each population, we randomly chose $2n$ haplotypes according to the haplotype frequencies and then randomly paired the haplotypes to form a population of n individuals with phase-known genotypes. The abilities of the two methods to reconstruct these haplotypes from the resulting data, ignoring phase information, were then evaluated. Twenty independent replicates for each population-locus combination were generated to compare the two haplotype reconstruction methods.

To estimate the haplotype frequencies, we implemented the EM algorithm in a computer program that analyzes the simulated data sets with the starting point of equal frequencies for every possible haplotype. We expect that any of the programs implementing the EM algorithm should yield similar results. Following Stephens et al. (2001), we specify the haplotype pair for an individual by choosing the most probable haplotype pair consistent with the individual's multisite genotype. The program developed by Stephens et al. (2001) was used to evaluate the performance of the PHASE method with

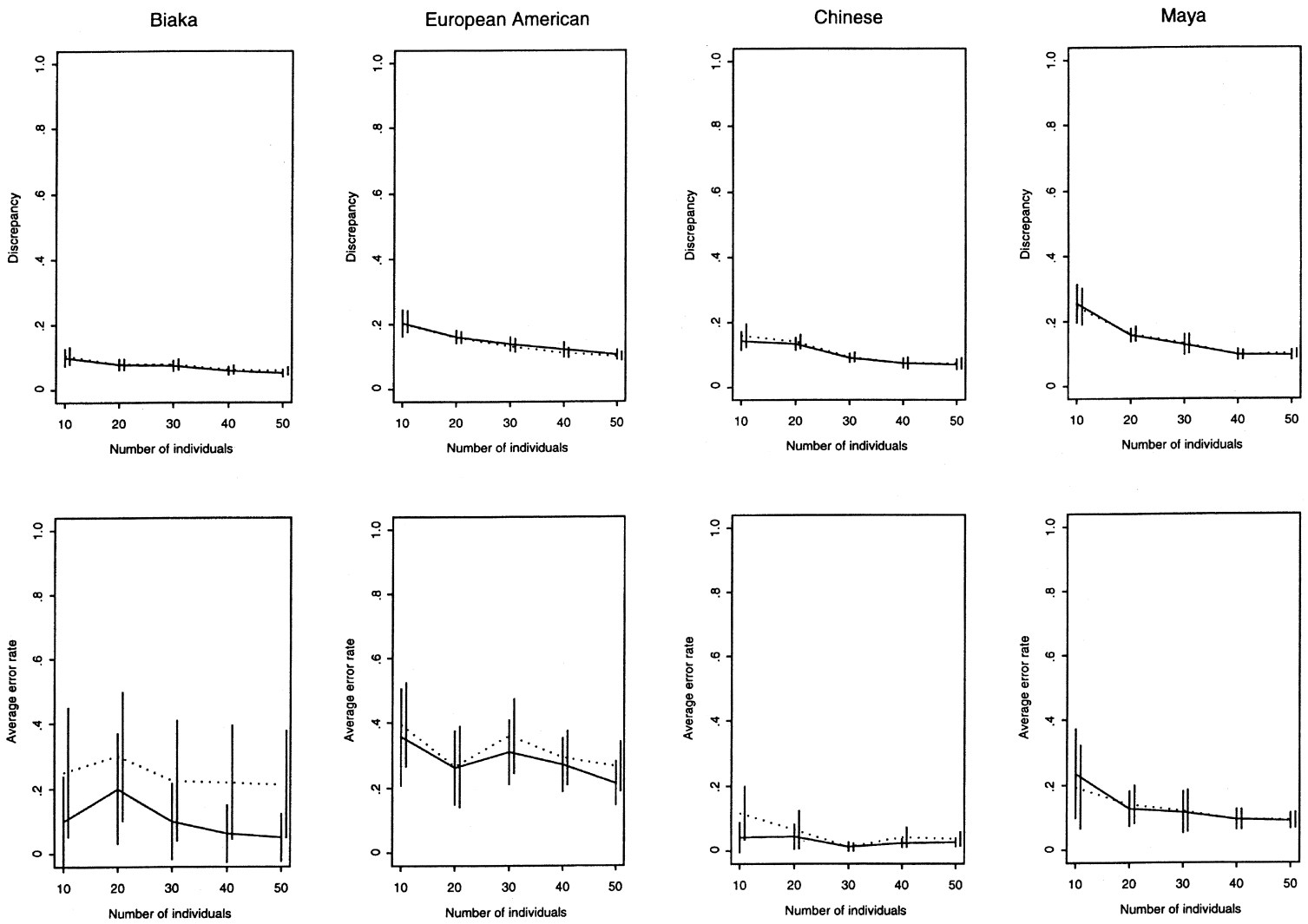

Figure 1 Comparisons between the EM method (*dotted lines*) and the PHASE method (*solid line*) for genotype data at the RET site, with three single-nucleotide polymorphisms (SNPs). For each scenario, we generate 20 independent data sets and, thus, each point represents an average of 20 simulated data sets. Vertical lines (left line for the PHASE method and right line for the EM method) show approximate 95% confidence intervals for the estimates (standard error = ± 2).

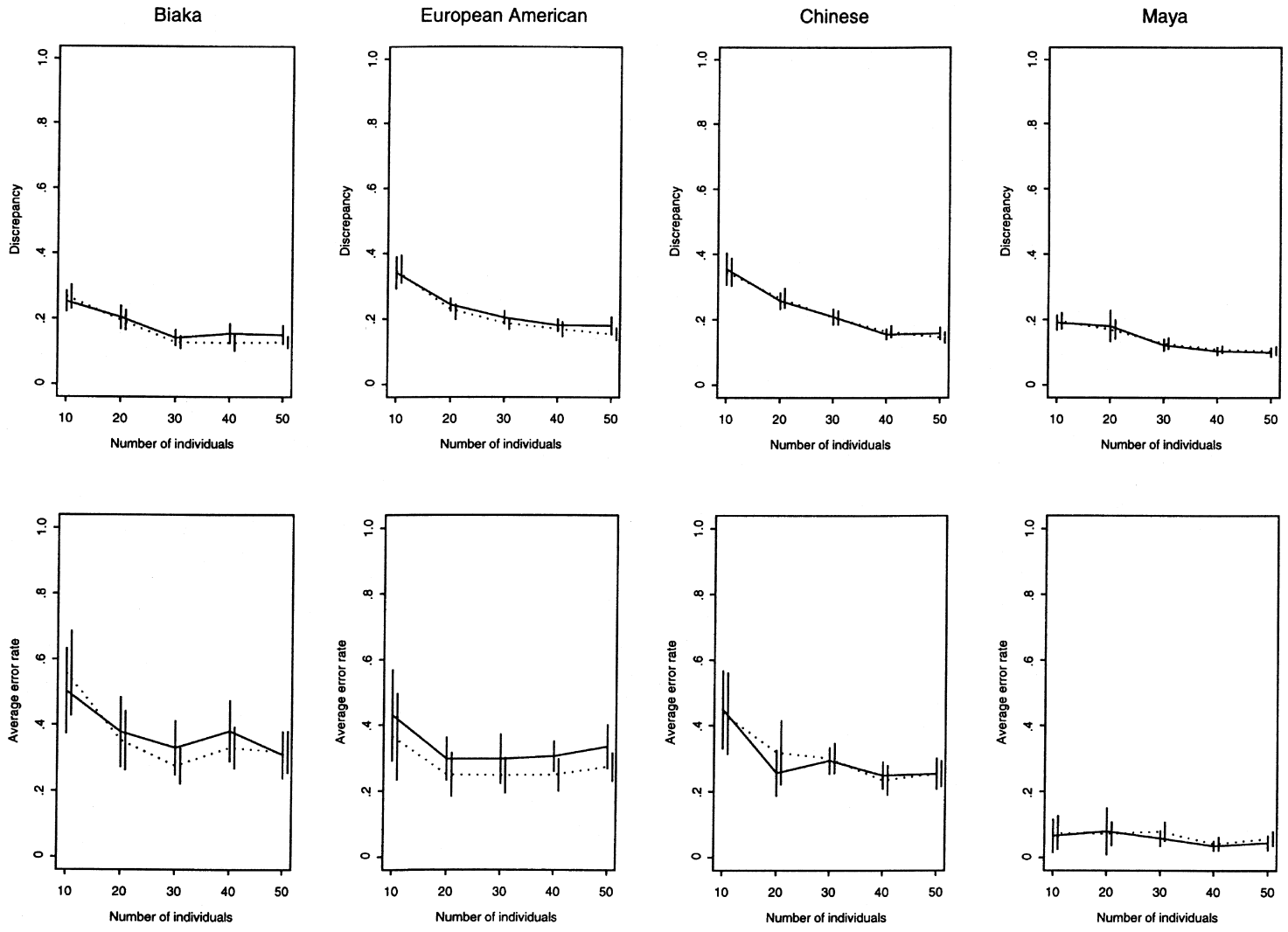

Figure 2 Comparisons between the EM method (*dotted lines*) and the PHASE method (*solid line*) for genotype data at the COMT site, with four SNPs. Conditions of each scenario, format of the graphs, and standard error are the same as those described in figure 1.

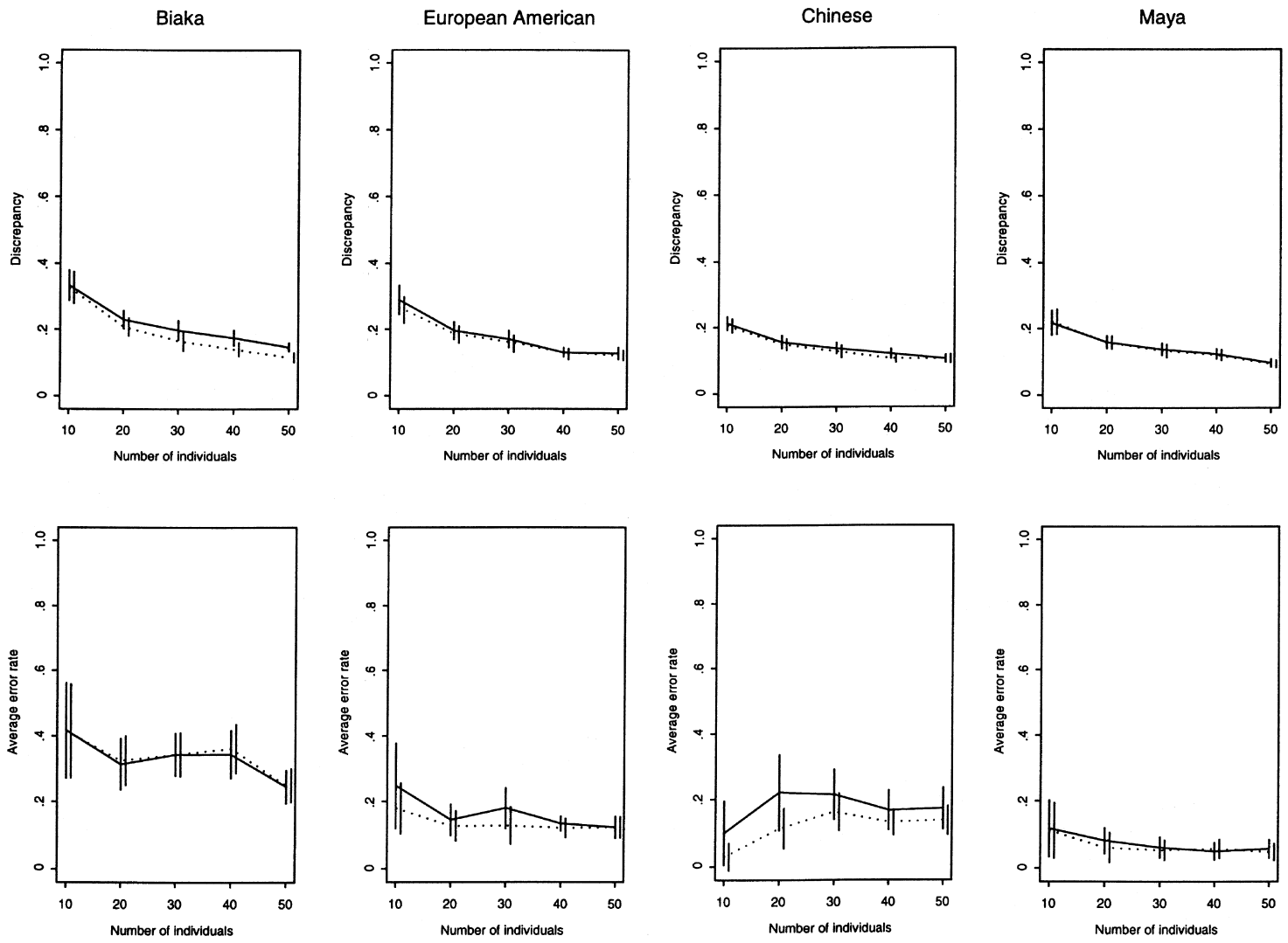

Figure 3 Comparisons between the EM method (*dotted lines*) and the PHASE method (*solid line*) for genotype data at the HOXB site with five SNPs. Conditions of each scenario, format of the graphs, and standard error are the same as those described in in figure 1.

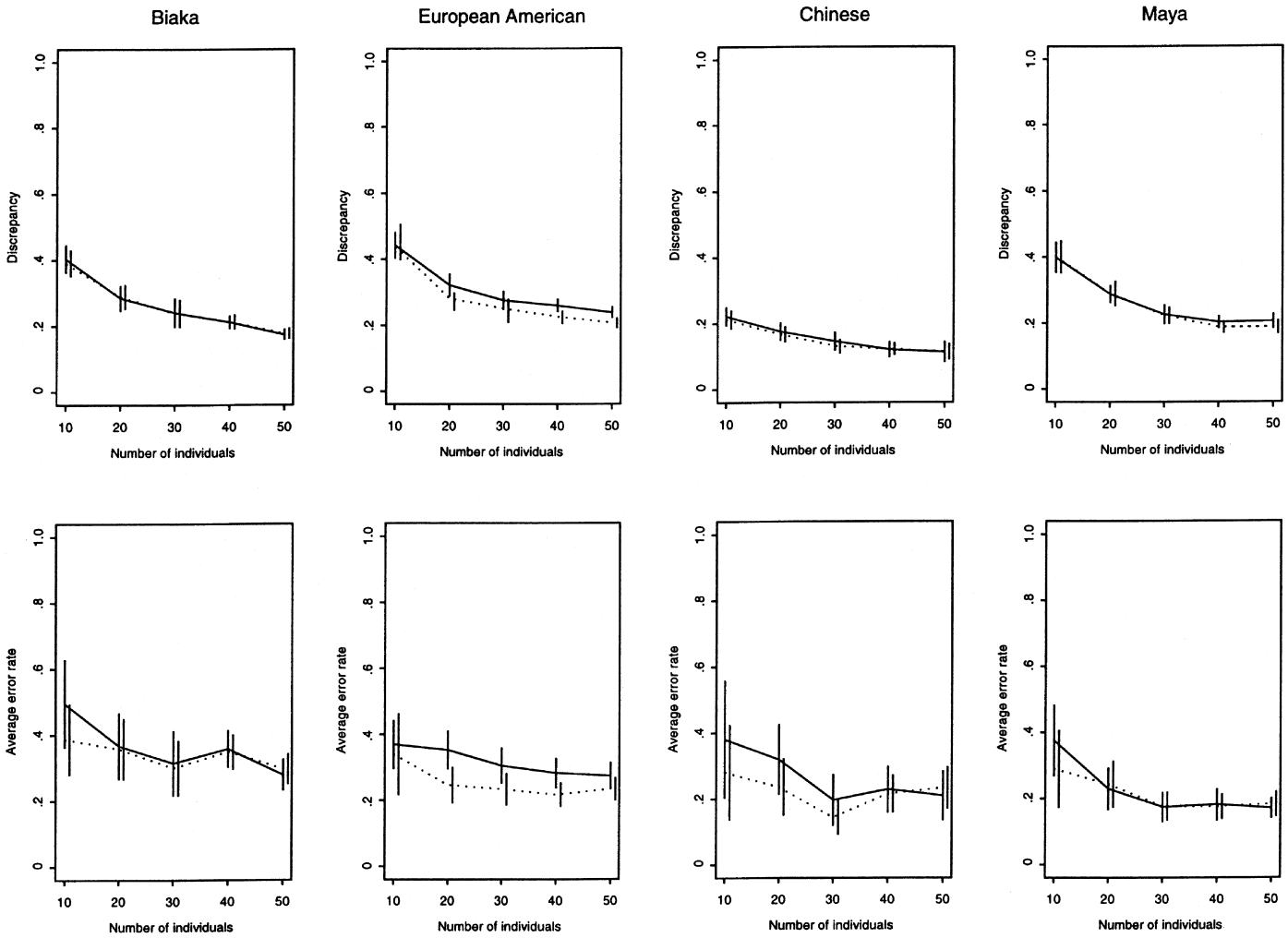

Figure 4 Comparisons between the EM method (*dotted lines*) and the PHASE method (*solid line*) for genotype data at the D4S10 site with six SNPs. Conditions of each scenario, format of the graphs, and standard error are the same as those described in figure 1.

Table 1

Comparisons between the EM Method and the PHASE Method, Using a Set of Phase-Known Data Sets at the CD4 Locus

Population	No. of Individuals	No. of Doubly Heterozygous Individuals	No. of Incorrectly Reconstructed Individuals		Discrepancy between True and Estimated Haplotype Frequencies	
			EM Method	PHASE Method	EM Method	PHASE Method
Biaka	53	8	3	3	.045	.057
Bantu	40	15	3	1	.089	.025
Herero	42	7	1	1	.024	.024
Mbuti	37	6	4	0	.086	0
Nama	32	5	2	2	.069	.069
Sekele	51	10	2	2	.036	.039
Wolof	46	13	3	3	.057	.065
Somali	24	5	0	0	0	0
Zu Wasi	44	5	2	2	.045	.034
Total	369	74	20	14	.05[a]	.034[a]

[a] Represents mean value for all nine populations.

the default parameter values in the Markov chain Monte Carlo simulations—that is, with 10,000 iterations, a thinning interval of 100, and a burn-in value of 10,000.

The comparison results for the four loci (each locus across four populations) are summarized in figures 1, 2, 3, and 4. The results show that, for almost all the cases we considered, the discrepancies between the estimated haplotype frequencies and the true haplotype frequencies are almost the same for the two methods. The average errors in haplotype reconstruction show slight differences across the four loci. The PHASE method gave better results than did the EM method, for the RET data sets with three polymorphisms; however, the EM method was better overall than the PHASE method for the other three loci—that is, for COMT, HOXB, and D4S10. The biggest difference between the results of the PHASE method and those of the EM method was found for the RET gene in the Biaka population. For this particular population/locus combination, only four of a possible total of eight haplotypes were inferred to be present, with the following haplotype frequencies: $P(000) = .089$, $P(001) = .747$, $P(011) = .029$, and $P(101) = .089$. In the above notation, the two alleles at each polymorphism are represented by 0 and 1, respectively. This situation seems optimal for a coalescent model, since each of the three uncommon haplotypes is one mutation away from the single very common haplotype. Samples drawn from this population would have few double heterozygotes, and a coalescent model would favor inferring the presence of haplotypes that are only one step away from the common haplotypes, rather than a haplotype two steps away. On the other hand, the EM algorithm will not add that bias. Despite the differences between the two methods, from the approximate 95% confidence intervals shown in the figures, we can see that there is no significant difference between these two methods, for most of the cases.

In our comparisons based on phase-known data sets, we used a subset of Tishkoff et al.'s (2000) CD4 genotype data, for nine populations (Biaka, South African Bantu, Herero, Mbuti, Sekele, Wolof, Somali, and Zu Wasi). Two markers, an *Alu* deletion polymorphism (2 alleles) and a microsatellite marker (12 alleles), were typed at CD4, and phases of doubly heterozygous individuals were determined molecularly (Tishkoff et al. 2000). The data and the results obtained by the EM method and the PHASE method are summarized in table 1.

There are a total of 74 doubly heterozygous individuals in nine populations. The error rates of the EM and the PHASE methods for haplotype reconstruction are 27% and 19%, respectively. The average discrepancies between haplotype estimates for the EM and PHASE methods are 5% and 3.4%, respectively. Therefore, across all of these nine populations, the PHASE method improved on the EM method by >30%; however, it can be seen from table 1 that the improvements did not come from across all of the populations. Instead, the two methods had identical performance in haplotype reconstruction for seven populations. In terms of average discrepancies, the PHASE method is better than the EM method for three populations, and the EM method is better than the PHASE method for three other populations. In the two populations for which the PHASE method outperformed the EM method—that is, the Bantu and Mbuti—the cause of the poorer performance of the EM method is the same as that for the simulation results based on empirical population haplotype frequency data. We note that even for the populations in which the two methods yielded the same number of in-

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.