Am. J. Hum. Genet. 67:518–522, 2000

Report

The Accuracy of Statistical Methods for Estimation of Haplotype Frequencies: An Example from the CD4 Locus

S. A. Tishkoff,[1,*] A. J. Pakstis,[1] G. Ruano,[2] and K. K. Kidd[1]

[1]Department of Genetics, Yale University School of Medicine, and [2]Genaissance Pharmaceuticals, Five Science Park, New Haven, CT

Haplotype analysis has become increasingly important for the study of human disease as well as for reconstruction of human population histories. Computer programs have been developed to estimate haplotype frequencies statistically from marker phenotypes in unrelated individuals. However, there currently are few empirical reports on the accuracy of statistical estimates that must infer linkage phase. We have analyzed haplotypes at the CD4 locus on chromosome 12 that consist of a short tandem-repeat polymorphism and an *Alu* insertion/deletion polymorphism located 9.8 kb apart, in 398 individuals from 10 geographically diverse sub-Saharan African populations. Haplotype frequency estimates obtained using gene counting based on molecularly haplotyped (phase-known) data were compared with haplotype frequency estimates obtained using the expectation-maximization algorithm. We show that the estimated frequencies of common haplotypes do not differ significantly with the use of phase-known versus phase-unknown data. However, rare haplotypes are occasionally miscalled when their presence/absence must be inferred. Thus, for those research questions for which the common haplotypes are most important, frequency estimates based on the phase-unknown marker-typing results from unrelated individuals will be sufficient. However, in cases where knowledge of rare haplotypes is critical, molecular haplotyping will be necessary to determine linkage phase unambiguously.

Haplotype and linkage disequilibrium analyses have become important tools for tracing population migration events (Wainscoat et al. 1986; Tishkoff et al. 1996; Harding et al. 1997; Kidd et al. 1998, 2000; Tishkoff et al. 1998) and for establishing founder effects for disease alleles (Hästbacka et al. 1992; Hoglund et al. 1995; Risch et al. 1995; Escamilla et al. 1996; Goldman et al. 1996; Stephens et al. 1998). Haplotypes consisting of one or more highly heterozygous short tandem-repeat polymorphisms (STRPs), sometimes in combination with biallelic flanking single-nucleotide polymorphisms (SNPs) or insertion/deletion polymorphisms, have become increasingly common. The high heterozygosity of STRPs and the increased diversity of haplotypes containing STRPs result in few homozygotes and, thus, few individuals with an unambiguous linkage phase. The primary data collected include the band sizes for each locus, and the phase of the alleles on the chromosomes is unknown for multiple-site heterozygotes. We use the terms "phase unknown" or "phenotype" to refer to an individual's marker-typing results in the absence of phase information, and we use the term "phase known" to refer to the individual's genetic constitution for the haplotyped system, including the linkage phase of the component marker alleles. When pedigree information is not available, linkage phase can either be established directly through molecular haplotyping or inferred probabilistically. Molecular haplotyping is the direct determination of linkage phase by generation and analysis of hemizygous templates from diploid genomic samples by use of allele-specific amplification (Michalatos-Beloin et al. 1996).

We are not aware of an empirical study demonstrating how closely the frequency estimates from phase-unknown data approximate those from gene-counting estimates based on phase-known data. The results of our own unpublished studies indicate that maximum like-

Received March 13, 2000; accepted for publication May 25, 2000; electronically published June 19, 2000.

Address for correspondence and reprints: Dr. Kenneth K. Kidd, Department of Genetics, Yale University School of Medicine, New Haven, CT 06520. E-mail: kidd@biomed.med.yale.edu

* Present affiliation: Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, PA.

0002-9297/2000/6702-0031$02.00

Table 1

Summary of Molecular Haplotyping

Population Sample	Sample Size	No. of Doubly Heterozygous Individuals (Phase Resolved)	Individuals Phase-Known Before Molecular Haplotyping (%)	No. of Different Kinds of Doubly Heterozygous Phenotypes[a]
Biaka	53	8 (8)	84.9	4
Mbuti	37	6 (6)	83.8	6
Sekele San	52	10 (10)	80.7	9
Zu/Wasi San	44	5 (5)	88.6	5
Nama	33	6 (5)	81.8	5
Kikuyu	22	3 (2)	86.4	3
Yoruba	17	4 (3)	76.5	3
Bantu	43	18 (16)	58.1	12
Woloff	48	17 (13)	64.6	13
Herero	49	14 (7)	71.4	11
Total	398	91 (75)	...	...

[a] Values are not counts of individuals.

lihood estimates based on samples with phase-unknown data could indicate the presence of haplotypes that are not actually present in the sample, and vice versa. We now have a large set of data that allows for systematic examination of this question for a nontrivial case. We have analyzed haplotypes at the CD4 locus (MIM 186940) on chromosome 12 that consist of two polymorphic markers located 9.8 kb apart: a biallelic polymorphism involving partial deletion of an *Alu* element (Edwards and Gibbs 1992) and a multiallelic pentanucleotide STRP (Edwards et al. 1991). The data on 398 individuals from 10 geographically diverse sub-Saharan African populations show little disequilibrium between these polymorphisms, and, because of the numerous STRP alleles, there are many different phase-unknown multiple heterozygotes (Tishkoff et al. 1996). These data provide an opportunity to compare the haplotype frequencies estimated, by use of gene counting based on molecularly haplotyped data (Michalatos-Beloin et al. 1996; Tishkoff et al. 1996), in the phase-known situation versus those haplotype frequencies estimated, by use of the expectation-maximization (EM) algorithm in the HAPLO program (Hawley and Kidd 1995), in the phase-unknown situation. (The source code for HAPLO, example data files, and documentation are available through the Kidd Lab Home Page Web site.) At least two programs that are similar to HAPLO and that use the EM algorithm have been independently developed by Excoffier and Slatkin (1995) and Long et al. (1995). All three programs have given us identical results on comparable analyses, and all three assume Hardy-Weinberg proportions. All 10 populations in the present study gave nonsignificant results for tests of Hardy-Weinberg ratios, for both the STR and *Alu* polymorphisms.

Using a primer specific for the *Alu* deletion allele (*Alu*[−]) and long-range PCR (Michalatos-Beloin et al. 1996; Tishkoff et al. 1996), we determined the phase-known genotypes for 75/91 doubly heterozygous individuals from a total of 398 individuals in the 10 sub-Saharan African populations sampled (table 1). Haplotypes were resolved completely in four populations and were 86%–97% resolved in the other six population samples. Incompletely resolved haplotypes occurred only in those individuals for which there were insufficient amounts of higher-molecular-weight DNA. In each of the populations for which linkage phase could be 100% determined, haplotype frequencies were estimated by direct gene counting. In each of the six populations in which a few haplotypes could not be unambiguously resolved, haplotype frequencies were estimated by incorporating the genotypes unambiguously defined by use of molecular haplotyping and allowing the EM algorithm in HAPLO to "assign" the few remaining ambiguous cases. The small proportion of phase-unknown multiple heterozygotes that remained could not alter the frequency estimates much from "pure" gene-counting estimates. Throughout this report, the data described above as well as the corresponding haplotype frequency estimates are referred to as "phase known." The data set on 398 individuals, which includes data for all 91 unresolved doubly heterozygous individuals along with the corresponding EM frequency estimates, is referred to as "phase unknown."

The haplotype frequency distributions estimated from the phase-known and phase-unknown data in each sample were not found to be statistically different (data not shown), on the basis of the Workman and Niswander heterogeneity test (1970). However, when phase is unknown, the EM algorithm will sometimes produce very low frequency estimates for haplotypes that are not actually present. For example, in the Mbuti sample, in

which 6/7 *Alu*(−) alleles exist as double heterozygotes, statistical inference generated six different *Alu*(−) haplotypes of roughly equal frequency (range .010–.023). In contrast, frequency estimates from the phase-known data revealed that only three of these haplotypes are actually present and that their frequencies (.053, .026, and .013) differ. Likewise, in some instances, haplotypes that were not predicted (i.e., those that had frequency estimates of zero) by the EM algorithm were demonstrated to be present by means of molecular haplotyping. For example, in the Bantu-speaking population, three haplotypes that were not predicted from phase-unknown data were shown to be present by molecular haplotyping.

When the standard errors for the haplotype frequency estimates were compared across the two data conditions, almost all the estimates were within 1.5 standard errors of each other (results not shown). This is consistent with the absence of an overall significant difference. The only estimates that differed by >2 standard errors were those for haplotypes that were absent in one of the two sets of estimates. We noticed a tendency for the larger absolute difference, as well as the standardized differences, to occur for haplotypes with small frequency estimates, and we developed a change index to show this graphically for all pairs of estimates. The change coefficient (C) incorporates both the direction and the percentage change in haplotype frequencies across the two information conditions. If we let H represent a haplotype frequency estimate based on phase-unknown data and if we let M represent a haplotype frequency estimate based on phase-known data, then C = (H−M)/Max[H,M], where Max[H,M] indicates the maximum value of H or M. C coefficients were computed for each possible CD4 haplotype in each of the sub-Saharan African populations. The value of C ranges from −1 to +1, with 0 indicating that the H and M estimates are the same. A negative value indicates that M > H, whereas M = −1 indicates that molecular haplotyping showed the presence of a haplotype that was assigned a zero frequency by the EM algorithm.

Figure 1 plots C on the *Y*-axis against the haplotype frequency level for each estimate (Max[H,M]) on the *X*-axis, for all possible haplotypes with nonzero frequency estimates determined by either analysis in any of the 10 population samples. The figure demonstrates that dramatic percentage changes ⩾90% occur only at the lowest haplotype frequencies (<5%), which, in this example, involved ~24/154 nonzero haplotype estimates across the 10 sub-Saharan African samples. Even the larger changes (range 30%–60%) occur only when the frequency estimates are <.10. Approximately two-thirds (68%) of the haplotype frequency estimates showed either no change or a small change

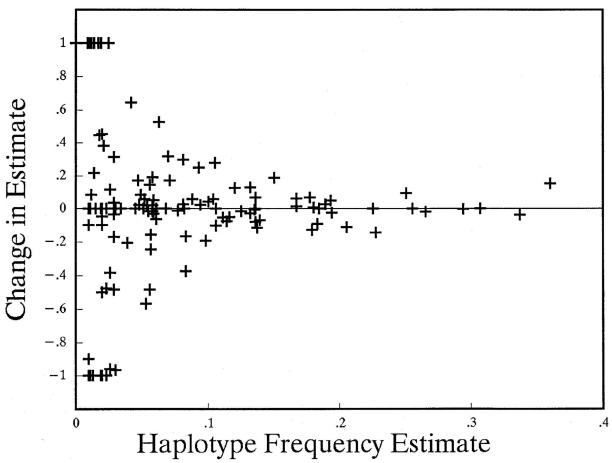

Figure 1 Haplotype frequency estimates with and without molecular haplotyping information are compared in samples from 10 African populations. The change in estimate value (denoted as "C" in the text) can range from −1 to +1, where −1 indicates a change from zero to a nonzero estimate after addition of molecular haplotyping information and where +1 indicates an estimate that changed to zero after the addition of molecular haplotyping information. The change value is plotted against the larger of the two frequency estimates.

(⩽15%). Some of the percentage changes that look as though they could be large actually involve a shift in very few chromosomes. For example, in figure 1, the rightmost point plotted (Max[H,M] ~.36; C 15%) involves a shift of only two chromosomes for the Yoruba 85bp/*Alu*(+) haplotype, since there are only 34 chromosomes in that sample. In that case, the jackknife standard error of the original phase-unknown estimate encompassed the estimate obtained after molecular haplotyping.

To assess, by use of phase-known and phase-unknown data, the effect of sample size on haplotype frequencies, C was plotted against haplotype frequency, as described above, with the sub-Saharan African populations grouped according to sample size: small (34–66 chromosomes), medium (74–88 chromosomes), and large (94–106 chromosomes) (data not shown). The limited effect of sample-size differences across the population samples studied is that, compared with larger sample sizes, the smaller sample sizes in this study (2*N* = 34–66) show a relative reduction in the number of low-frequency haplotypes. Consequently, fewer dramatic changes in frequency occur across the estimation conditions in the smaller samples, since the dramatic changes in estimates are concentrated at rare (<.05) haplotype frequencies. As the sample size grows, there are more opportunities to observe rare haplotypes, almost all of which will be in heterozygotes. Overall, however, sample size did not have a large effect on the haplotype

frequency estimates comparing phase-known and phase-unknown results.

We also compared the standard errors for the phase-unknown estimates with the standard errors for the estimates incorporating molecular haplotyping (data not shown). As expected, in virtually all cases, the standard errors were lower for the phase-known estimates reflecting the additional information provided. Approximately half of the haplotypes showed large decreases in the standard errors of the estimates when phase-known data were used, and almost all of the remaining haplotypes had only small decreases in the standard errors of the frequency estimates. Of 154 standard-error estimates, 16 were slightly larger for the phase-known estimates, reflecting small increments in the frequency of the haplotypes involved; 8 of the 16 involved haplotype frequencies changing from zero (phase unknown) to a small value (phase known).

The example presented is not a "worst-case" scenario but neither is it trivial. The 10 African populations studied for CD4 polymorphisms display little or no linkage disequilibrium between the two markers, and a significant proportion (~23%) of the sampled individuals are double heterozygotes for whom no single phenotype (combination of marker-typing results) is overwhelmingly frequent (table 1). It is easy to see, in contrasting situations with strong disequilibrium, that there can be a high proportion of phase-unknown multiple heterozygotes that are relatively easy to resolve because only a few combinations of marker typings predominate; molecular haplotyping may add little information in such individuals. For instance, for the CD4 locus, the data from a study by Tishkoff et al. (1996) indicate that the European population has a higher proportion of multiple heterozygotes (~42%), but these divide roughly equally between two common doubly heterozygous genotypes generated by the three common haplotypes—85bp/*Alu*(+), 110bp/*Alu*(+), and 90bp/*Alu*(−)—accounting for 92.6% of all chromosomes in the European population. Because of this nearly complete association, the standard error on haplotype frequencies estimated by use of HAPLO is low. The very strong linkage disequilibrium in non-African populations clearly compensates for the high overall proportion of phase-unknown multiple heterozygotes.

In the Taiwanese Han sample of a previous study (Lu et al. 1996), a similar example occurs at the DRD2 locus, where the single most common phenotype (18/46 individuals) is a triple heterozygote but where the next two most common (10 and 5 individuals) phenotypes are "opposite" triple homozygotes. The four remaining phenotypes are mostly unambiguous with respect to the underlying genotypes. Common sense—and the EM algorithm—will assign each of the triple heterozygotes to a single genotype, and one can anticipate little increase in accuracy from molecular haplotyping. However, in situations in which most individuals are multiply heterozygous and no single phenotype is common, as when two STRP loci are in equilibrium, most haplotypes will be uncommon, and molecular haplotyping might result in a very different profile, compared with that given by the EM algorithm. The CD4 example is intermediate between these two extremes.

These examples highlight the importance of both the relative frequencies and the numbers of different phase-unknown multiply heterozygous phenotypes, rather than simply the total proportion of phase-unknown multiple heterozygotes, in assessment of the accuracy of haplotype estimation. The existence of many different phenotypes at roughly similar frequencies implies that many different haplotypes occur at low frequency and that, hence, greater error and uncertainty occur in the estimation of haplotype frequencies. In contrast, the presence of a small number of multiply heterozygous phenotypes at proportionately high frequencies implies that some individual haplotypes exist at high frequencies, and the estimation of those haplotypes' frequencies will be accomplished with greater accuracy.

Our findings for CD4 haplotypes show that the estimated frequencies of common haplotypes do not differ significantly using phase-known versus phase-unknown data. For research in which the common haplotypes are most important, frequency estimates based on the phase-unknown marker-typing data (phenotypes) of unrelated individuals will be sufficient. However, accurate identification of rare haplotypes may be essential for some areas of research, including detection of low levels of population admixture and definition of rare mutation-bearing haplotypes in the absence of family material. Also, when multiple polymorphic sites display little or no disequilibrium, a large percentage of the chromosomes in a population may occur as uncommon or rare haplotypes. In such cases, the overall level of uncertainty from phase-unknown data will be larger. These comparisons demonstrate the usefulness of molecular haplotyping for determining linkage phase from typing results on highly polymorphic markers as well as for obtaining more-accurate haplotype frequency estimates when pedigree information is unavailable.

Acknowledgments

We thank the reviewers for their comments. This project was supported in part by United States Public Health Service grants GM57672, MH39239, MH30929, AA09379, and National Science Foundation DBS-9208197 (to K.K.K.) and by NSF SBIR grant DMI-9461102 (to G.R.).

Electronic-Database Information

The accession number and URLs for data in this article are as follows:

Kidd Lab Home Page, http://info.med.yale.edu/genetics/kkidd (for source code of HAPLO and related software)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for CD4 [MIM 186940])

References

Edwards MC, Clemens PR, Tristan M, Pizzuti A, Gibbs RA (1991) Pentanucleotide repeat length polymorphism at the human CD4 locus. Nucleic Acids Res 19:4791

Edwards MC, Gibbs RA (1992) A human dimorphism resulting from loss of an *Alu*. Genomics 14:590–597

Escamilla MA, Spesny M, Reus VI, Gallegos A, Meza L, Molina J, Sandkuijl LA, et al (1996) Use of linkage disequilibrium approaches to mapping genes for bipolar disorder in the Costa Rican population. Am J Med Genet 67:244–253

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

Goldman A, Krause A, Ramsay M, Jenkins T (1996) Founder effect and prevalence of myotonic dystrophy in South Africans: molecular studies. Am J Hum Genet 59:445–452

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin S, et al (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772–789

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

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

Hoglund P, Sistonen P, Norio R, Holmberg C, Dimberg A, Gustavson KH, de la Chapelle A, et al (1995) Fine mapping of the congenital chloride diarrhea gene by linkage disequilibrium. Am J Hum Genet 57:95–102

Kidd JR, Pakstis AJ, Zhao H, Lu R-B, Okonofua FE, Odunsi A, Grigorenko E, et al (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH*, in a global representation of populations. Am J Hum Genet 66:1882–1899

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, et al (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227

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

Lu R-B, Ko HC, Chang FM, Castiglione CM, Schoolfield G, Pakstis AJ, Kidd JR, et al (1996) No association between alcoholism and multiple polymorphisms at the dopamine D2 receptor gene (DRD2) in three distinct Taiwanese populations. Biol Psychiatry 39:419–429

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

Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, et al (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9:152–159

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, et al (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507–1515

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Cheung K, Kidd JR, Bonne-Tamir B, et al (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Goldman A, Calafell F, Speed WC, Bonne-Tamir B, Kidd JR, Pakstis AJ, et al (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Wainscoat JS, Hill AV, Boyce AL, Flint J, Hernandez M, Thein SL, Old JM, et al (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature 319:491–493

Workman PL, Niswander JD (1970) Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Am J Hum Genet 22:24–49