Letter to the Editor

Am. J. Hum. Genet. 67:1626–1627, 2000

Efficiency of Estimation of Haplotype Frequencies: Use of Marker Phenotypes of Unrelated Individuals versus Counting of Phase-Known Gametes

To the Editor:
Tishkoff et al. (2000) compare two-locus haplotype-frequency estimates based on marker phenotypes with estimates based on counting of phase-known gametes. They conclude that marker phenotypes of unrelated individuals are adequate for estimation of the frequencies of common haplotypes but that counting of phase-known gametes is preferable when estimates of the frequency of rare haplotypes are required.

As Tishkoff et al. note, the choice between these two strategies for estimation of haplotype frequencies is of practical importance in various situations. It is therefore of interest to examine the statistical efficiency of haplotype-frequency estimation based on marker phenotypes, compared with that based on counting of phase-known gametes. We can derive this by comparing the *observed information* (curvature of the log-likelihood function) at the maximum-likelihood value of the haplotype frequency with the *complete information* (information that we would have if phase were known for all gametes in the sample). We consider a sample of unrelated individuals typed at two loci, A and B. Without loss of generality when estimating the frequency of a single haplotype, we number the allelic states that constitute the haplotype under study, as "A_1" and "B_1," grouping together all other allelic states at locus A as "A_2" and all other allelic states at locus B as "B_2." We use p_{ij} for the population frequency of the haplotype, A_iB_j. We define the efficiency of haplotype-frequency estimation based on marker phenotypes as the ratio of the expected observed information to the expected complete information. For a two-locus haplotype, this is given by the following expression (derived in the Appendix):

$$\frac{\text{Expected observed information}}{\text{Expected complete information}} = \frac{p_{11}p_{22}(p_{11}+p_{22})+p_{12}p_{21}(p_{12}+p_{21})}{p_{11}p_{22}+p_{12}p_{21}} \,. \quad (1)$$

If there is no allelic association, then $p_{11}p_{22} = p_{12}p_{21}$, and haplotype-frequency estimation based on marker phenotypes is 50% efficient. If there is allelic association and all four haplotypes have frequency $\geqslant$1%, then the efficiency of estimation based on marker phenotypes can be in the range 39%–98%, depending on the allele frequencies and the strength of allelic association. It remains to be determined how far these conclusions can be generalized to estimation of haplotype frequencies at three or more loci.

This relationship provides a theoretical basis for some of the empirical results reported by Tishkoff et al. (2000): when allelic association is weak or absent, as in the African populations that they studied, two-locus haplotype–frequency estimates based on marker phenotypes will have approximately twice the variance of estimates based on counting of phase-known gametes; when allelic association is strong, as in the non-African populations that they studied, determination of phase will not necessarily add much extra information about haplotype frequencies. However, Tishkoff et al.'s conclusion that haplotype-frequency estimation based on marker phenotypes performs less well for rare haplotypes than for common haplotypes should be qualified. When the frequency of the haplotype under study (A_1B_1) is 1%, the efficiency of haplotype-frequency estimation based on marker phenotypes can be either as low as 39% (when the frequencies of alleles A_1 and B_1 are both .16 and these alleles are inversely associated) or close to 100% (when haplotypes A_1B_2 and A_2B_1 predominate in the population). It is true that, even when haplotype A_1B_1 is absent in the sample of individuals studied, estimation based on marker phenotypes can maximize the likelihood at a nonzero value for the population haplotype frequency p_{11}. However, one would not simply rely on the point estimate $\hat{p}_{11}$ to infer that the true value of p_{11} is >0.

The efficiency of haplotype-frequency estimation based on marker phenotypes has practical implications for the design of genetic-association studies in which haplotype frequencies will be compared between cases and controls. We may ask whether it is worth typing the parents or offspring of cases and controls in order to be able to assign gametic phase. From equation (1) we can infer that the information about two-locus haplotype frequencies from a sample of 100 unrelated individuals with no missing marker phenotypes will be equivalent to that obtained from a sample of 78–200 phase-known gametes, depending on the allele frequencies and allelic association in the population. To obtain a sample of 150 phase-known gametes by typing the parent-offspring pairs, it would be necessary to type >50 parent-offspring pairs, because, even when parent-offspring pairs are typed, not all haplotypes can be unambiguously inferred. Thus, for estimation of two-locus haplotype frequencies in controls, the two strategies do not differ much in terms of the information obtained for a given genotyping workload. For haplotype-frequency estimation in cases, typing a parent or offspring of each case in order to determine phase does not contribute extra gametes to the sample on the basis of which haplotype frequencies in cases are estimated, and it is therefore more economical to type a sample of unrelated cases than to type a sample of case-offspring or case-parent pairs. However, haplotype-frequency estimation based on marker phenotypes relies on the assumption that the haplotypes are in Hardy-Weinberg equilibrium. For cases, this assumption will hold only if the haplotype risk ratios fit a multiplicative model (Clayton 1999).

PAUL M. MCKEIGUE
Department of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine
London

Appendix

Suppose that we type a sample of N individuals at two diallelic loci, A and B, where the allele frequencies are (p_A,q_A) and (p_B,q_B). If we observe only the marker phenotypes, then the observed data vector consists of the number n of double heterozygotes and the counts y_{ij} of haplotypes A_iB_j in other marker phenotypes. The missing data vector consists of the number x of double heterozygotes whose alleles are in coupling phase. When allele frequencies are held at their maximum-likelihood values, estimation of the haplotype frequencies is equivalent to estimation of the disequilibrium coefficient D. The complete-data log-likelihood is

$$\begin{aligned}&(y_{11}+x)\log(p_Ap_B+D)\\&\quad+(y_{12}+n-x)\log(p_Aq_B-D)\\&\quad+(y_{21}+n-x)\log(q_Ap_B-D)\\&\quad+(y_{22}+x)\log(q_Aq_B+D)\ .\end{aligned}$$

The score (gradient of the complete-data log-likelihood as a function of D) is

$$\frac{y_{11}+x}{p_Ap_B+D}-\frac{y_{12}+n-x}{p_Aq_B-D}-\frac{y_{21}+n-x}{q_Ap_B-D}+\frac{y_{22}+x}{q_Aq_B+D}\ .$$

The *complete information* (curvature of the complete-data log-likelihood) is

$$\frac{y_{11}+x}{(p_Ap_B+D)^2}+\frac{y_{12}+n-x}{(p_Aq_B-D)^2}+\frac{y_{21}+n-x}{(q_Ap_B-D)^2}+\frac{y_{22}+x}{(q_Aq_B+D)^2}\ .$$

For simplicity of notation, the haplotype frequencies $[p_Ap_B+D,p_Aq_B-D,q_Ap_B-D,q_Aq_B+D]$ are written below as $[p_{11},p_{12},p_{21},p_{22}]$.

The variance of the score over the posterior distribution of the missing data, given the observed data, is the *missing information* (Little and Rubin 1987). This is equal to

$$\left(\frac{1}{p_{11}}+\frac{1}{p_{12}}+\frac{1}{p_{21}}+\frac{1}{p_{22}}\right)^2\mathrm{V}(x)\ ,$$

where $\mathrm{V}(x)$ is the variance of x over its posterior distribution, given by $n\prod(1-\prod)$, where $\prod=p_{11}p_{22}/(p_{11}p_{22}+p_{12}p_{21})$. The expectation of n is $(2p_{11}p_{22}+2p_{12}p_{21})N$, and therefore the expected missing information is

$$\left(\frac{1}{p_{11}}+\frac{1}{p_{12}}+\frac{1}{p_{21}}+\frac{1}{p_{22}}\right)^2\frac{p_{11}p_{12}p_{21}p_{22}}{p_{11}p_{22}+p_{12}p_{21}}2N\ .$$

The expected complete information is

$$\left(\frac{1}{p_{11}}+\frac{1}{p_{12}}+\frac{1}{p_{21}}+\frac{1}{p_{22}}\right)2N\ .$$

The observed information is calculated by subtracting the missing information from the complete information (Louis 1982). Dividing the expected observed information by the expected complete information yields equation (1).

References

Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177

Little RJA, Rubin DB (1987) Statistical analysis with missing data. John Wiley & Sons, New York

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc Series B 44:226–232

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

Address for correspondence and reprints: Dr. Paul M. McKeigue, Department of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom. E-mail: paul.mckeigue@lshtm.ac.uk

0002-9297/2001/6706-0033$02.00