Am. J. Hum. Genet. 78:728–729, 2006

Impact of Haplotype-Frequency Estimation Error on Test Statistics in Association Studies

To the Editor:

Curtis and Sham (2006 [in this issue]) discuss the impact of haplotype-frequency estimation error on several statistics in case-control association studies. Of course, it is worth investigating how the error of haplotype-frequency estimation affects the results of association studies. However, to assess systematically the impact of haplotype-frequency estimation errors on association studies is not a trivial task. Here, we only consider some simple cases, to respond to the letter from Curtis and Sham (2006).

Consider the simple case of two loci, each with two alleles, yielding four haplotypes with frequencies f_1, f_2, f_3, and f_4. Consider an example in which sample size $n = 500$, $f_1 = f$, $f_2 = 2f$, $f_3 = 3f$, and $f_4 = 1 - f_1 - f_2 - f_3$. Under the null hypothesis of equal haplotype frequencies in both cases and controls, figure 1 shows

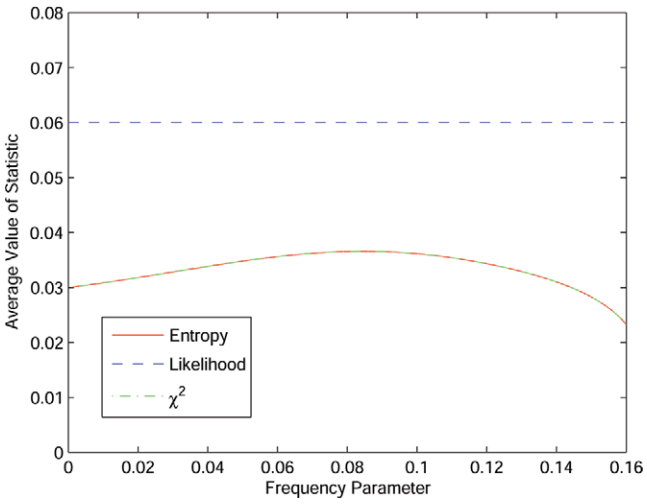

Figure 1 Asymptotic average values for the χ^2, entropy-based, and likelihood-ratio–based heterogeneity test statistics as a function of the frequency parameter f, under the assumptions $n = 500$, $f_1 = f$, $f_2 = 2f$, $f_3 = 3f$, and $f_4 = 1 - f_1 - f_2 - f_3$.

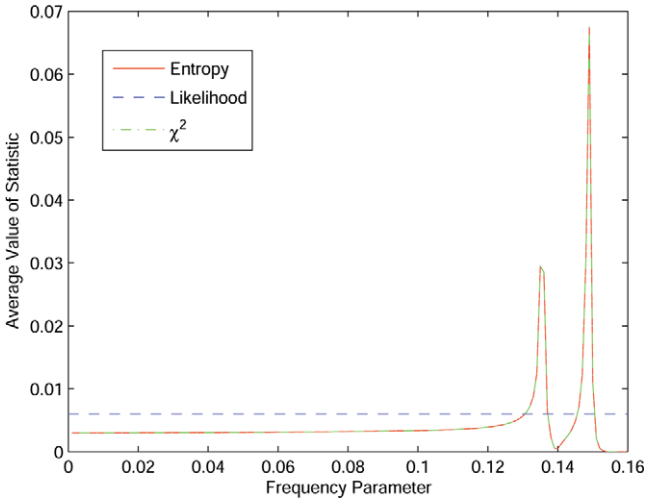

Figure 2 Asymptotic average values for the χ^2, entropy-based, and likelihood-ratio–based heterogeneity test statistics as a function of the frequency parameter f, under the assumptions $n = 500$, $f_1 = 4f$, $f_2 = 1.5f$, $f_3 = 1.5f$, and $f_4 = 1 - f_1 - f_2 - f_3$.

the asymptotic average values for the χ^2, entropy-based, and likelihood-ratio–based heterogeneity test statistics as a function of the parameter f. In theory, under the null hypothesis, the values of these three statistics should be equal to zero. However, because of sampling error and haplotype-frequency estimation error, these three statistics are not actually equal to zero. Figure 1 shows that, on average, the χ^2 statistic and the entropy-based statistic are always smaller than the likelihood-ratio–based heterogeneity statistic. This implies that, for these haplotype frequencies, the χ^2 statistic and the entropy-based statistic should have smaller type I error rates than that of the likelihood-ratio–based heterogeneity test statistic. To analyze the example given by Curtis and Sham (2006), we created the graph in figure 2. Figure 2 shows the values of these three statistics when $n = 500$, $f_1 = 4f$, $f_2 = 1.5f$, $f_3 = 1.5f$, and $f_4 = 1 - f_1 - f_2 - f_3$. We can see that, in most ranges of the parameter f, the average values of the χ^2 statistic and the entropy-based statistic are smaller than that of the likelihood-ratio–based heterogeneity test statistic. However, when the frequencies of the four haplotypes are close to 0.5, 0.2, 0.2, and 0.1—which are the haplotype frequencies used

to generate table 1 in the letter from Curtis and Sham (2006)—the average values of the χ^2 statistic and the entropy-based statistic are much larger than that of the likelihood-ratio–based heterogeneity test statistic. The range of the haplotype frequencies leading to this result is not large. Both figures 1 and 2 demonstrate that the average values of the χ^2 statistic and the entropy-based statistic are similar and that, in most cases, the average values of both the χ^2 statistic and the entropy-based statistic are smaller than that of the likelihood-ratio–based heterogeneity test statistic.

If the covariance matrices of the moment estimates of haplotype frequencies in the standard χ^2 statistic and the entropy-based statistic are replaced with the covariance matrix of maximum-likelihood estimates of the haplotype frequencies, the average values of the three statistics will be asymptotically the same. Therefore, if we were to use the covariance matrix from the maximum-likelihood estimates, any differences among the three test statistics would be, on average, small.

In summary, (1) we formulate the entropy-based statistic in terms of the estimated frequencies, not counts as implemented in the simulation performed by Curtis and Sham (2006); (2) the estimation error for haplotype frequencies will have an impact on the test statistic under the null hypothesis, and the magnitude of this effect will depend on the haplotype frequencies; and (3) asymptotically, the impact of the haplotype-frequency estimation error on the standard χ^2 statistic and the entropy-based statistic is smaller than the impact on the likelihood-ratio–based heterogeneity test statistic. Therefore, the claim that type I error rates of the heterogeneity test are always much smaller than those of the standard χ^2 statistic and the entropy-based statistic is incorrect. The effects of the haplotype-frequency estimation errors on type I error rates of a test are complex and should be investigated by both theoretical analysis and intensive simulation studies over large parameter spaces, not just over a small range of haplotype frequencies.

JINYING ZHAO, ERIC BOERWINKLE, AND MOMIAO XIONG

Human Genetics Center
University of Texas Health Science Center at Houston
Houston

Reference

Curtis D, Sham PC (2006) Estimated haplotype counts from case-control samples cannot be treated as observed counts. Am J Hum Genet 78:729–730 (in this issue)

Address for correspondence and reprints: Dr. Momiao Xiong, Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX 77030. E-mail: Momiao.Xiong@uth.tmc.edu

0002-9297/2006/7804-0020$15.00

Am. J. Hum. Genet. 78:729–730, 2006

Estimated Haplotype Counts from Case-Control Samples Cannot Be Treated as Observed Counts

To the Editor:

Although the entropy-based method described by Zhao et al. (2005) provides a sensitive way to detect kinds of departure from a random distribution of haplotype counts between two data sets, we cannot see how it can be applied in practice to case-control samples. This is because tests that treat haplotype counts estimated from unphased data as if they were actually observed haplotypes are inherently anticonservative.

To illustrate in principle why this is, let us consider a sample in which all subjects happen to be doubly heterozygous at two loci, with genotype Aa/Bb. The maximum-likelihood estimates for the haplotype frequencies do not consist, as one might think intuitively, of each possible haplotype having frequency 0.25. Instead, there are two equally likely solutions: that haplotypes AB and ab each occur with frequency 0.5 or that, conversely, haplotypes Ab and aB both have frequency 0.5. (With N subjects, the solution for four haplotypes has likelihood 0.25^{2N}, whereas the solution for just two haplotypes has likelihood 0.5^{2N}.) If a few other genotypes are added to the data set, they will push the solution one way or the other. For example, if the sample consists of a mixture of cases and controls, and one case has genotype AA/Bb and one control has genotype aa/Bb, then the estimated haplotype frequencies will suggest that almost all cases have haplotypes AB and ab, whereas almost all controls have haplotypes Ab and aB. Although such an extreme example would not occur in practice, it is important to understand that maximum-likelihood estimation of haplotype frequencies favors solutions containing a small number of different haplotypes. This implies that, when frequencies of multilocus haplotypes are estimated separately in cases and controls, small random effects can produce quite large, apparently notable differences. In a real situation, one might estimate, for instance, that a particular haplotype occurred in a small percentage of cases but never in controls, leading to the possibly erroneous deduction that this indicates the presence of a pathogenic mutation.

To determine whether haplotype frequencies differ sig-