Re: The Use of Inferred Haplotypes in Downstream Analysis

To the Editor: In a letter published in the March 2007 issue of the *Journal,* Lin and Huang[1] described some potential pitfalls in haplotype analysis when only unphased genotyped data are available. In particular, they relate the problems associated with picking the most probable haplotype from a distribution of potential haplotypes to use in risk analysis, replacing the unobserved haplotypes with their "best estimates." This particular form of "single imputation" can indeed be dangerous. However, another form of single imputation exists that is far more reliable, as has been shown in a series of articles describing analytical results and a series of simulation experiments.[2,3] This "correct" single-imputation method is the expectation-substitution method first described for haplotype analysis by Zaykin et al.,[4] but which has analogues in many parts of biostatistical analysis dealing with measurement errors.[5,6]

Consider a model in which the number of copies carried of a certain haplotype, *h,* is related to the odds of disease in a log-linear fashion. The expectation-substitution method works as follows: since the number of copies of the risk haplotype carried, n_h, is unknown for most individuals when only genotype data are available, one replaces this unknown quantity with its conditional expectation, $E(n_h|G)$, where G is the observed set of (unphased) markers, and the expectation is generally computed under the assumption that the haplotype frequencies are known and under the assumption of Hardy-Weinberg equilibrium. This expectation (which generally takes noninteger values 0–2 and thus is definitely not equivalent to using the most likely value of n_h, which must take values 0, 1, or 2) is then used as if it were the true value of n_h, with no other allowance for the uncertainty of the estimation of n_h. In particular, we perform score tests and likelihood-ratio tests of the null hypothesis (that h is unrelated to risk) as if n_h were known for all subjects. This method is not limited to log-linear penetrance models and can be extended to dominant, recessive, or codominant models by calculating the expected values of appropriate diplotype codings.[2] It has been used in a large number of recent analyses of haplotype-specific risk in candidate-gene studies.[7–10]

The following issues arise in the application of this method:

1. Because haplotype frequencies must be estimated from a finite amount of genotype data, the calculation of the expectation is itself uncertain, and this uncertainty is being neglected. What effect does this have?
2. Because odds-ratio (OR) models are not strictly linear in n_h, the expected OR as a function of n_h is not equal to the OR function applied to $E(n_h)$. Is this an important problem?
3. Doesn't case-control sampling distort haplotype frequencies? Should the expectation be applied separately to the cases and controls, or should the cases and controls be combined to compute the expectation?

Regarding issue 1, it turns out[11] that the score test arising from the expectation method is both valid and asymptotically fully efficient for testing the null hypothesis of no haplotype-specific risk associated with haplotype h against the alternative (that disease odds are log-linear in n_h). Thus, for score tests under the null hypothesis, there is no need to account for the uncertainty of the estimates of the haplotype frequencies on which the calculation of $E(n_h)$ is based. Similarly, under the null hypothesis, case-control sampling does not distort the distribution of the risk haplotype, so the calculation of the expectation is not affected. This means that we do not need to account for case-control sampling either when testing the null hypothesis.

Simulations under the alternative hypothesis[2] tend to show that, under the alternative hypothesis (that disease is associated with haplotype count), the expectation-substitution method also gives quite reasonable estimates and confidence limits for the value of the risk parameter (log-OR per copy of h) in most practical settings. Here, we present the results of two simulation experiments. We considered estimating the association between five-SNP haplotypes and disease risk in two situations: low haplotype diversity (table 1) and high haplotype diversity (table 2). For the first situation, only seven haplotypes were present: 00000, 00001, 00010, 00011, 00100, 01000, and 10000 with frequencies 0.35, 0.15, 0.15, 0.05, 0.1, 0.1, and 0.1, respectively ($R_h^2 > 0.7$ for all haplotypes, where R_h^2 is the squared correlation of expected versus true haplotype counts[12]). For the second situation, all 32 possible haplotypes were equally likely. This is something of a worst-case scenario for haplotype association analysis, both because of the high ambiguity in haplotypes given the genotypes ($R_h^2 \approx 0.43$ for all haplotypes) and because all haplotypes are relatively uncommon (frequency <0.04). We compared the performance of logistic regression, using the true haplotypes, the most likely haplotype pair given the genotype data, the expectation-substitution method, and the prospective likelihood implemented in the haplo.glm function of the R package haplo.stats.[13] This last approach is similar in spirit to that advocated by Lin and Huang,[1] since it maximizes a likelihood that integrates over the missing haplotype-phase information. (We chose haplo.glm primarily for computational simplicity. This allowed us to perform simulations, analysis, archiving, and

Table 1. Comparison of Haplotype-Association Methods under Low Haplotype Diversity

	Relative Risk = 1.0					Relative Risk = 1.5					Relative Risk = 3.0				
Method	Bias	MSE	Var(β)	E(Var)	Cover	Bias	MSE	Var(β)	E(Var)	Cover	Bias	MSE	Var(β)	E(Var)	Cover
True haplotypes	.005	.016	.016	.016	.960	.009	.016	.015	.015	.938	.005	.013	.013	.015	.958
Expectation substitution	.006	.021	.021	.018	.940	.008	.017	.017	.017	.968	−.002	.015	.015	.017	.968
Maximum likelihood	.006	.021	.021	.018	.940	.009	.017	.017	.017	.968	.000	.015	.015	.017	.968
Most-likely haplotypes	.004	.016	.016	.015	.948	−.026	.015	.014	.014	.946	−.069	.017	.013	.014	.919

NOTE.—Based on 500 simulated studies of 600 cases and 600 controls. Genotypes were drawn conditional on disease status, with the assumption of the haplotype structure described in the text and a log-linear model of disease risk, with risk haplotype 00001 and baseline disease probability of 1%. For each method, log-ORs for six haplotypes (the 00000 haplotype was set as reference) were estimated jointly. Bias = average estimated log-OR minus the true log-OR; MSE = mean squared error in estimate for risk-haplotype log-OR; Var(β) = variance of estimated log-OR; E(Var) = mean of estimated parameter estimate variance; cover = empirical coverage of nominal 95% CI.

summary in one statistical package, R. Although HAPSTAT, the graphical user interface implementation of the maximum-likelihood methods developed by Lin et al.,[14–16] is quite user-friendly and suitable for the analysis of single data sets, it does not appear to be easily automated for the analysis of hundreds of simulated data sets.)

As can be seen from table 1, under low haplotype diversity, the expectation-substitution and the maximum-likelihood method give indistinguishable results. In fact, the average euclidean distance between the vectors of haplotype log-ORs for the two methods is <0.007 for all the situations presented in table 1, and the average distance between the haplotype frequency vectors was <0.001. Neither method has noticeable bias, and both have appropriate coverage for ORs of 1.5 and 3.0. The most-likely-haplotype method is slightly biased toward the null.

On the other hand, both the expectation-substitution and maximum-likelihood methods show evidence of bias under high haplotype diversity (table 2). The expectation-substitution method shows modest bias toward the null, whereas the maximum-likelihood method shows stronger bias away from the null. Moreover, the nominal 95% CIs for the expectation-substitution method have appropriate coverage, whereas those for the maximum-likelihood method are far too small, and the average estimated variance for the estimated log-OR is noticeably smaller than the observed variance in the maximum-likelihood estimates. We hypothesize that this is due to the relatively large number of highly colinear parameters (64) that must be jointly estimated from data from 600 cases and 600 controls. We emphasize that the high-diversity situation should not often arise in practice, because haplotype association analyses are generally restricted to regions of low haplotype diversity (appropriately so, in our opinion), but it is interesting that even in this situation the expectation-substitution method shows only modest bias and retains appropriate coverage, whereas the maximum-likelihood method performs poorly, perhaps because of numerical difficulties.

Although we used the expectation-maximization algorithm to estimate haplotype frequencies and posterior haplotype probabilities conditional on individual genotypes, one could also use more-sophisticated algorithms, such as those implemented in PHASE.[17] PHASE has been shown to provide more-accurate estimates of haplotype frequencies than does the expectation-maximization algorithm,[18] in part because it models mutation and recombination processes, whereas the expectation-maximization algorithm only assumes Hardy-Weinberg equilibrium. The fact that the expectation-substitution approach allows the user to choose from a range of haplotype-frequency–estimation algorithms is a potential advantage over the maximum-likelihood approach, although, for most situations where haplotype association analysis is applied—small numbers of SNPs in high linkage disequilibrium over short distances—we anticipate that the difference between haplotype-frequency estimates from PHASE and from the expectation-maximization algorithm will be quite small.

Our conclusions based on this type of simulation are

Table 2. Comparison of Haplotype-Association Methods under High Haplotype Diversity

	Relative Risk = 1.0					Relative Risk = 1.5					Relative Risk = 3.0				
Method	Bias	MSE	Var(β)	E(Var)	Cover	Bias	MSE	Var(β)	E(Var)	Cover	Bias	MSE	Var(β)	E(Var)	Cover
True haplotypes	−.003	.111	.111	.113	.950	.036	.102	.101	.104	.952	.039	.086	.085	.098	.960
Expectation substitution	−.007	.683	.684	.587	.952	−.019	.485	.485	.508	.970	−.046	.335	.336	.414	.980
Maximum likelihood	−.027	1.241	1.242	.205	.590	.069	.986	.984	.166	.618	.213	.886	.849	.142	.551
Most-likely haplotypes	−.008	.288	.288	.257	.950	−.271	.301	.228	.196	.884	−.697	.634	.150	.163	.449

NOTE.—Based on 500 simulated studies of 600 cases and 600 controls. Genotypes were drawn conditional on disease status, with the assumption that all 32 possible five-SNP haplotypes are equally likely and with a log-linear model for disease risk, with risk haplotype 00001 and baseline disease probability of 1%. For each method, log-ORs for six haplotypes (the 00000 haplotype was set as reference) were estimated jointly. Bias was calculated relative to the true ORs for the risk haplotype. Bias = average estimated log-OR minus the true log-OR; MSE = mean squared error in estimate for risk-haplotype log-OR; Var(β) = variance of estimated log-OR; E(Var) = mean of estimated parameter estimate variance; cover = empirical coverage of nominal 95% CI.

that the expectation-substitution method provides very reliable inference (correct type I error rates under the null hypothesis), good power under alternatives, and little bias either in overall estimates or in confidence limits. It appears to be that only when the true ORs become extremely large do some problems occur with the method, and, frankly, from an epidemiological perspective, we should be so lucky as to have very many association studies with this problem!

Peter Kraft and Daniel O. Stram

Acknowledgments

P.K. was supported by National Cancer Institute (NCI) grants U01 CA098233 and P01 CA08796. D.O.S. was supported by National Human Genome Research Institute grant GM58897 and NCI grant 5P30 ES07048.

References

1. Lin DY, Huang BE (2007) The use of inferred haplotypes in downstream analyses. Am J Hum Genet 80:577–579
2. Kraft P, Cox DG, Paynter RA, Hunter D, De Vivo I (2005) Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet Epidemiol 28:261–272
3. Cordell HJ (2006) Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. Genet Epidemiol 30:259–275
4. Zaykin D, Westfall P, Young S, Karnoub M, Wagner M, Ehm M (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91
5. Thomas D, Stram D, Dwyer J (1993) Exposure measurement error: influence on exposure-disease relationships and methods of correction. Ann Public Health 14:69–93
6. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C (2006) Measurement error in nonlinear models: a modern perspective, 2nd ed. Chapman and Hall, New York
7. Haiman CA, Stram DO, Pike MC, Kolonel LN, Burtt NP, Altshuler D, Hirschhorn J, Henderson BE (2003) A comprehensive haplotype analysis of CYP19 and breast cancer risk: the Multiethnic Cohort. Hum Mol Genet 12:2679–2692
8. Cox DG, Kraft P, Hankinson SE, Hunter DJ (2005) Haplotype analysis of common variants in the BRCA1 gene and risk of sporadic breast cancer. Breast Cancer Res 7:R171–R175
9. Zhai R, Gong MN, Zhou W, Thompson TB, Kraft P, Su L, Christiani DC (2007) Genotypes and haplotypes of VEGF gene are associated with higher mortality and lower VEGF plasma levels in patients with ARDS. Thorax 62:718–722
10. Tamimi RM, Cox DG, Kraft P, Pollak MN, Haiman CA, Cheng I, Freedman ML, Hankinson SE, Hunter DJ, Colditz GA (2007) Common genetic variation in IGF1, IGFBP-1, and IGFBP-3 in relation to mammographic density: a cross-sectional study. Breast Cancer Res 9:R18
11. Xie R, Stram DO (2005) Asymptotic equivalence between two score tests for haplotype-specific risk in general linear models. Genet Epidemiol 29:166–170
12. Stram D, Haiman C, Hirschhorn J, Altshuler D, Kolonel L, Henderson B, Pike M (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using as preliminary sample of unrelated subjects with an example from the multiethnic cohort study. Hum Hered 55:27–36
13. Lake S, Lyon H, Tantisira K, Silverman E, Weiss S, Laird N, Schaid D (2003) Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55:56–65
14. Zeng D, Lin DY, Avery CL, North KE, Bray MS (2006) Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. Biostatistics 7:486–502
15. Lin D, Zeng D (2006) Likelihood-based inference on haplotype effects in genetic association studies (with discussion). J Am Stat Assoc 101:89–118
16. Lin DY, Zeng D, Millikan R (2005) Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol 29:299–312
17. Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462
18. Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, et al (2006) A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78:437–450

From the Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston (P.K.); and Division of Biostatistics, Keck School of Medicine, University of Southern California, Los Angeles (D.O.S.)

Address for correspondence and reprints: Dr. Peter Kraft, Harvard School of Public Health, Building 2 Room 207, 665 Huntington Avenue, Boston, MA 02115. E-mail: pkraft@hsph.harvard.edu

Am. J. Hum. Genet. 2007;81:863. © 2007 by The American Society of Human Genetics. All rights reserved.
0002-9297/2007/8104-0031$15.00
DOI: 10.1086/522899

Reply to Peter Kraft and Daniel O. Stram

To the Editor: The main purpose of our original letter[1] was to show that the common practice of using the most probable haplotype in association analysis can be dangerous. We are glad that Kraft and Stram share this view and provide numerical support.[2(in this issue)] We agree with them that the expectation-substitution method is generally preferable to the use of the most probable haplotype. Because it ignores the phenotype information and the case-control sampling in the imputation, however, this method can still yield biased and inefficient analysis of association. In our original letter,[1] we reported the power estimates of 62%, 49%, 42%, and 50% for detecting the effects of haplotypes D, F, G, and H, respectively, in a simulation study mimicking that of French et al.[3] The corresponding power estimates for the expectation-substitution method are 56%, 42%, 36%, and 42%. Thus, the expectation-substitution method is considerably less powerful than the maximum-likelihood method.

The simulation results shown in table 2 of the letter by