The Use of Inferred Haplotypes in Downstream Analyses

To the Editor: In the March 2006 issue of the *Journal,* Marchini et al.[1] provided a comprehensive description of phasing algorithms for inference of individual haplotypes from unphased genotype data. They stated that an unresolved question is "whether and, if so, how best to use inferred haplotypes in downstream analyses."[1(p.448)] The question is important because knowledge of individual haplotypes is rarely an end in itself. We offer our perspective on this issue, particularly in the context of (case-control) association studies.

Phase ambiguity is a kind of missing data, and use of inferred haplotypes in downstream analyses is a form of imputation. The voluminous statistical literature on missing data casts light on the potential pitfalls of imputation. In the words of Dempster and Rubin,[2(p.8)]

> The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial bias.

As pointed out by Marchini et al.,[1] all the phasing algorithms assume Hardy-Weinberg equilibrium (HWE). Even when the general population is in HWE, the case sample and the pooled case-control sample may not be.[3] Thus, the phasing algorithms may produce biased estimation of haplotype distributions with case-control data. The influence of departures from HWE on estimation accuracy depends on the directionality of the disequilibrium.[4] The phasing algorithms do not acknowledge the selective-sampling feature of the case-control design and thus may produce biased results. Also, the phasing algorithms do not take into account phenotype, which is potentially informative about phase.

The common practice of assigning the most likely diplotype (i.e., the pair of haplotypes with the highest posterior probability) to each individual is intrinsically biased because the most likely diplotype is not necessarily the true diplotype. Consider the simple situation of two SNPs, with the minor and major alleles coded as 1 and 0, respectively, at each SNP site. The genotype is defined as the number of minor alleles at the two SNP sites. Haplotype ambiguity arises if and only if an individual is doubly heterozygous—that is, has the 11 genotype. Both the (10,01) and (00,11) diplotypes produce the 11 genotype. There is obviously a problem if all doubly heterozygous individuals are assigned the more likely (i.e., the more common) of the two diplotypes, especially when the frequency of the less common diplotype is similar to (although lower than) that of the more common diplotype.

When causal haplotypes exist, the phasing algorithms may incorrectly assign causal haplotypes to individuals without causal haplotypes or may reconstruct causal haplotypes as noncausal haplotypes. Consequently, treatment of inferred haplotypes as true haplotypes in downstream association analyses tends to attenuate the estimated haplotype effects and to reduce the power for detecting causal variants. Incorrect haplotype assignments may also induce spurious association for noncausal haplotypes and thus increase false-positive results.

For illustration, we consider the diplotype distribution from a hypothetical case-control study shown in the top part of table 1. With diplotype D as the reference, the estimated odds ratios (ORs) for diplotypes A, B, and C are 3, 1, and 1, respectively. Assume that, for both cases and controls, 20% of the individuals that truly have diplotype A are incorrectly assigned diplotype B, and another 20% are incorrectly assigned diplotype D, yielding the misclassified distribution shown in the bottom part of table 1. Then, the estimated OR for diplotype A is reduced from 3 to 2.3; for diplotype B, it is increased from 1 to 1.2; and, for diplotype C, it is reduced from 1 to 0.8. This example demonstrates that treatment of inferred haplotypes as true haplotypes may bias the estimated effects of causal haplotypes downward and may also bias the estimated effects of noncausal haplotypes away from the null value in either direction. The distortions may be more profound if the misclassification rates differ between cases and controls.

Several simulation studies[5–8] showed that imputation can yield substantial bias of estimated genetic effects, poor coverage of confidence intervals, and significant inflation of type I error, especially when the effects are large and the phase uncertainty is high. A recent article by French et al.[8] reported the bias of the estimated log ORs in the range of −0.49 to 0.22, an actual type I error of 18% at the 5% nominal significance level, and coverage of <40% for 95% CIs. Indeed, when the estimator is biased, the coverage of the associated 95% CIs will decrease toward 0% and the type I error will increase toward 100% as the sample size increases.

In recent years, researchers[3,9–12] have developed maximum-likelihood methods to properly account for phase uncertainty in association analyses. This approach involves maximizing the observed-data likelihood with respect to all relevant parameters (including haplotype frequencies and disease risks) simultaneously. The maximum-likelihood estimators for haplotype effects are unbiased and statistically efficient, which implies that maximum likelihood is the most powerful among all valid methods.[12] A question naturally arises as to how much more powerful maximum likelihood is compared with im-

Table 1. Effects of Incorrectly Assigned Haplotypes on Risk Estimates

Type of Haplotype and Measure	Diplotype			
	A	B	C	D
True haplotypes:				
No. of cases	500	100	200	200
No. of controls	250	150	300	300
OR	3.0	1.0	1.0	…
Inferred haplotypes:				
No. of cases	300	200	200	300
No. of controls	150	200	300	350
OR	2.3	1.2	0.8	…

putation. To answer this question, we conducted a few simulation studies.

In our first simulation study, we considered the type 1 angiotensin receptor gene (*AGTR1*) studied by French et al.[8] There were 12 SNPs with nine "common" haplotypes (see table I in the work of French et al.[8]). We generated case-control data under the third model given in their table III, but we used more-moderate ORs of 2.5, 2, 1.5, and 2 for haplotypes D, F, G, and H, respectively. We assigned disease status under the additive mode of inheritance, such that the disease prevalence was ~2%, and we selected 800 subjects, with 3 control subjects per case subject. On the basis of 10,000 simulated data sets with 2% randomly missing SNPs, the power of the maximum-likelihood method[11] to detect the effects of haplotypes D, F, G, and H was estimated at 62%, 49%, 42%, and 50%, respectively, at the nominal significance level of 1%, compared with power of 50%, 39%, 24%, and 32%, respectively, for the imputation method with the PHASE (v2.1) algorithm. The maximum-likelihood and imputation methods yielded type I error of 1% and 2%, respectively, for the null haplotype E at the 1% nominal significance level. In this study, ~75% of individuals had unambiguous diplotypes, and ~82% had highest posterior probabilities >0.75.

Our second simulation study mimicked the two-locus model Mul3 of Cordell.[7] We assumed that haplotypes 01 and 10 have ORs of 1.2 and 1.4 in reference to haplotypes 00 and 11, with additive mode of inheritance, and we tested whether locus 2 has an effect, while allowing an effect at locus 1. On the basis of 10,000 simulated data sets of 1,000 cases and 1,000 controls with 10% randomly missing genotypes, we obtained power values of 65%, 40%, and 17% at the nominal significance levels of 5%, 1%, and 0.1%, respectively, for the maximum-likelihood method, compared with 41%, 20%, and 6% power for the imputation method.

As pointed out by a referee, the phasing algorithms reviewed by Marchini et al.[1] are often used to phase large regions, so it would be interesting to assess the performance of the imputation method in testing for haplotype-disease association on a small set of SNPs that is phased within a larger genomic context. To this end, we generated 100 SNPs according to the allele frequencies and pairwise linkage disequilibrium (LD) coefficients of the first 100 SNPs on chromosome 18 of the CEU sample in the HapMap genomewide data, and we performed haplotype analysis on SNPs 60–64. The most common haplotypes of the five SNPs are 00000, 00001, 00010, 00100, 00101, 01101, 10000, 10001, 10010, 10100, and 10101, with frequencies of 4.6%, 8.8%, 11.0%, 7.4%, 7.2%, 7.0%, 6.6%, 6.8%, 8.6%, 7.4%, and 8.4%, respectively. We assumed that the disease risk was influenced by haplotype 00000 only, with an OR of 3 under the additive mode of inheritance. We set the overall disease prevalence to ~5% and selected 300 cases and 300 controls. We assessed the haplotype-disease association for those 5 SNPs, which were phased together with the other 95 SNPs by the PHASE algorithm. It was not computationally feasible to phase 600 subjects altogether for the 100 SNPs. Thus, we randomly divided the 600 subjects into six groups of 50 cases and 50 controls. (We found that phasing cases and controls together provided much better control of type I error than did phasing cases and controls separately.) We simulated 1,000 data sets with 2% randomly missing SNP values. We found that, at the nominal significance level of 1%, the imputation method had 60% power to detect the causal haplotype 00000 and had type I error of 5%, 3%, 4%, and 7% for null haplotypes 00001, 00010, 00100, and 10000, respectively, whereas the maximum-likelihood method had 72% power to detect the causal haplotype and had type I error close to the nominal level for all null haplotypes. The maximum-likelihood estimates had little bias, whereas the imputation method produced bias of −0.33, 0.27, 0.21, 0.26, and 0.30 for the log ORs of haplotypes 00000, 00001, 00010, 00100, and 10000, respectively.

In the above study, the LD among the five SNPs was not particularly strong (table 2). In a related study, we considered SNPs 95–99, which had very high LD (table 3). The most common haplotypes of SNPs 95–99 are 00000, 00001, 01000, 01001, 01100, 01111, 10000, and 10001, with frequencies of 39.7%, 20.8%, 2%, 1.3%, 1.8%, 13.8%, 12.9%, and 5.4%, respectively. We assumed that 10001 is the causal haplotype with an OR of 2.5 under the additive mode of inheritance. The rest of the simulation setup was the same as in the previous simulation study. The imputation method had 83% power to detect the causal haplotype and had type I error of 2% and 4% for null haplotypes 00001 and 10000 at the nominal significance level

Table 2. Standardized LD Coefficients (*D′*) for SNPs 60–64 on Chromosome 18 of the HapMap CEU Sample

SNP	*D′* for SNP			
	61	62	63	64
60	1.0	.86	.28	.68
61	…	.86	1.0	.84
62	…	…	.55	.73
63	…	…	…	.51

Table 3. Standardized LD Coefficients (*D*′) for SNPs 95–99 on Chromosome 18 of the HapMap CEU Sample

	D′ for SNP			
SNP	96	97	98	99
95	1.0	1.0	1.0	.96
96	…	.83	.95	.94
97	…	…	.95	.77
98	…	…	…	.94

of 1% and produced bias of −0.15, 0.12, and 0.14 for the log ORs of haplotypes 10001, 00001, and 10000, respectively. On the other hand, the maximum-likelihood method had 92% power to detect the causal haplotype and provided accurate control of type I error and unbiased estimates of haplotype effects.

Our studies obviously do not encompass all possible scenarios. Thus, the results do not imply that imputation is always bad, but rather that it can be considerably less powerful than maximum likelihood while providing biased estimates of genetic effects and poor control of type I error in practical situations. The problems tend to be more severe when there is greater uncertainty in reconstructed haplotypes.

Our simulation studies were focused on single imputation, which is the most common practice. Some alternative procedures have been proposed, including multiple imputation, expectation substitution, and weighted logistic regression.[6–8] These procedures are not theoretically valid either (for many of the reasons mentioned above) and may perform poorly. In particular, the versions of multiple imputation that have been proposed are improper because they fail to account for phenotype and case-control sampling. Proper multiple imputation would provide a good approximation to maximum likelihood. In short, no method can be more powerful than maximum likelihood while providing the same control of type I error, although some methods may approximate maximum likelihood well under certain circumstances. We recommend that maximum likelihood be generally adopted for analyses of haplotype-disease associations.

A major appeal of imputation is that standard statistical software can be used to perform the desired association analyses, once individual haplotypes are inferred by a phasing algorithm. The extent to which maximum likelihood can be used in association analyses depends critically on the availability of specialized software. Several groups have developed computer programs for maximum-likelihood methods. We recently posted a user-friendly software interface called "HAPSTAT." This software provides maximum-likelihood procedures for estimating and testing haplotype effects and haplotype-environment interactions under a wide variety of disease models.

D. Y. Lin and B. E. Huang

Web Resource

The URL for data presented herein is as follows:

HAPSTAT, http://www.bios.unc.edu/~lin/software/

References

1. 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
2. Dempster AP, Rubin DB (1983) Introduction. In: Madow WG, Olkin I, Rubin DB (eds) Incomplete data in sample surveys, volume 2: theory and bibliography. Academic Press, New York, pp 3–10
3. Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329
4. Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for bialleli loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959
5. Morris AP, Whittaker JC, Balding DJ (2004) Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide–polymorphism genotype data. Am J Hum Genet 74:945–953
6. 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
7. 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
8. French B, Lumley T, Monks SA, Rice KM, Hindorff LA, Reiner AP, Psaty BM (2006) Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol 30:485–494
9. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434
10. Spinka C, Carroll RJ, Chatterjee N (2005) Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol 29:108–127
11. 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
12. Lin DY, Zeng D (2006) Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc 101:89–118

From the Department of Biostatistics, University of North Carolina, Chapel Hill

Address for correspondence and reprints: Dr. Danyu Lin, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB 7420, Chapel Hill, NC 27599-7420. E-mail: lin@bios.unc.edu

Am. J. Hum. Genet. 2007;80:577. © 2007 by The American Society of Human Genetics. All rights reserved.
0002-9297/2007/8003-0021$15.00
DOI: 10.1086/512201