Letters to the Editor

Am. J. Hum. Genet. 71:992–995, 2002

Caution on Pedigree Haplotype Inference with Software That Assumes Linkage Equilibrium

To the Editor:

For the purpose of gene discovery, haplotypes can provide significant information. Haplotypes can increase the information for linkage with a set of markers, each of which on its own may provide weak information, and haplotypes can be used to identify genotype errors, through identification of double recombination over short chromosomal regions. In addition, the inference of haplotypes from pedigree data has been advocated as a method to refine positional cloning of disease genes (Sobel et al. 1996). Haplotypes can be used to identify chromosomal regions that are common among affected persons in isolated populations, because these subjects may have inherited a conserved chromosomal segment from a common founder. The implicit basis of fine mapping by haplotypes is linkage disequilibrium (LD) between alleles of the underlying susceptibility locus and at least one of the marker loci. This is due to the small number of recombinations of the disease and marker loci, which preserves the founder haplotype among affected individuals. Although it is possible for the alleles of different marker loci to be in linkage equilibrium among themselves yet be in disequilibrium with the disease locus, this will most likely occur only when the disease-causing allele is much younger than the origin of the marker alleles. Hence, if closely spaced markers are useful for haplotype fine mapping, it is reasonable to assume that the markers themselves are in LD. However, most commonly used software packages that can be used for the inference of haplotypes for pedigree members assume linkage equilibrium among the markers. Despite this assumption, it is not unusual for investigators to proceed with haplotype fine mapping by inferring haplotypes by the use of software that assumes no LD. Even the documentation for GENEHUNTER 1.0 states that "Haplotypes can be invaluable tools … in searching for shared genomic regions of distantly related affected individuals and indicating linkage disequilibrium between markers…." However, this practice can be misleading, as we discovered in our analysis of the association of prostate cancer (PC) with haplotypes composed of three marker loci within the *HPC1* gene. This example is provided to illustrate the impact that LD can have on haplotype inference.

The *HPC1/RNASEL* gene on chromosome 1q25 was recently identified as a candidate gene for hereditary PC (Carpten et al. 2002). To test for potential gene associations and increased risk for disease, three missense polymorphisms (Ile97Leu, Arg462Gln, and Glu541Asp) were genotyped in 432 patients with familial PC and in 470 population-based control subjects (Wang et al. 2002). These three loci, with linear order 97, 462, then 541, are in close proximity (loci 97 and 462 separated by 1.1 kb; loci 462 and 541 separated by 3.2 kb), so our intent was to compare the frequencies of haplotypes composed of these three loci between the patients with familial PC and the unrelated population-based control subjects. Haplotype frequencies among the unrelated control individuals were estimated by an "EM algorithm" (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995), as implemented in S-PLUS software (Schaid et al. 2002), which enumerates all possible haplotypes for each subject and uses the combined data from all control subjects (including those with and without haplotype ambiguities) to estimate the haplotype frequencies. This algorithm assumes that the haplotypes are randomly paired for each subject, which implies that each of the loci is in Hardy-Weinberg equilibrium. Hardy-Weinberg equilibrium was tested for each of the three loci, among the controls, and only locus 541 suggested some departure ($P = .06$). This locus tended to have fewer heterozygotes than expected (44.9% observed vs. 49.2% expected), which does not tend to bias the EM algorithm for the estimation of haplotype frequencies (Fallin and Schork 2000).

The 435 patients with familial PC came from 178 pedigrees (1–7 patients genotyped per pedigree). Haplotype frequencies among the familial patients were estimated by first inferring the most likely haplotypes for each pedigree member and then computing frequencies among the inferred haplotypes. To infer haplotypes, we used the program GENEHUNTER (Kruglyak et al. 1996), which is based on determination of the maximum-likelihood set of inheritance vectors that explain the data, under the assumption of linkage equilibrium of the marker loci. The marker allele frequencies were based on the control in-

dividuals, although the frequencies of the "1" alleles were similar between case and control subjects (1.9% vs. 1.3% for Ile97Leu; 30.6% vs. 37.2% for Arg462Gln; 47.5% vs. 43.5% for Glu541Asp). We allowed minimal chance of recombination (recombination fraction = 0.001) between adjacent markers. The resulting haplotype frequencies are presented in table 1. The most remarkable difference between GENEHUNTER haplotype frequencies for cases versus those for controls estimated by the EM algorithm is for haplotype 5, which has a frequency of 11.8% among cases but did not occur among controls.

Because the method used to estimate haplotype frequencies differed between cases and controls, we repeated our analysis but randomly sampled one case per family, and then used the EM algorithm to compute haplotype frequencies among the cases. The results from this analysis (table 1) show that the haplotype frequencies did not differ between cases and controls and that the main difference is caused by the use of GENEHUNTER for pedigrees versus the EM algorithm for a sample of unrelated cases. Even when we repeated this analysis, making a new random selection of cases, similar results were found. Why would GENEHUNTER, which uses all the pedigree data, produce such different haplotype frequencies from those produced by the EM algorithm, which uses only one random case per family? Two factors influenced our results. First, the markers are in strong LD, yet GENEHUNTER assumes that they are in equilibrium. The EM algorithm does not assume linkage equilibrium (in fact, LD can be a benefit). The strength of LD is shown in table 1 by the large discrepancy between estimated haplotype frequencies among controls when the EM algorithm was used and the estimated haplotype frequencies when linkage equilibrium was assumed. Furthermore, the frequencies of the inferred haplotypes among cases are close to the expected frequencies among the random sample of cases if there were no LD (see table 1). Second, the order in which the alleles are coded in the input file to GENEHUNTER can determine the inferred haplotype. A simple example in our data nicely illustrates the problem. If we consider an affected pair of brothers who do not have parental genotypes available and who are both heterozygous for each of the three loci, there are four possible pairs of haplotypes per man. These possible haplotype pairs are listed in table 2, along with the posterior probabilities of the haplotype pairs. The posterior probability is the probability of a particular pair of haplotypes, given the observed marker data, and was computed when the EM algorithm was applied to the randomly selected sample of one patient per family. This posterior probability allows for LD, and we also present (table 2) the posterior probabilities when no LD is assumed (i.e., equal posterior probabilities).

For this example, the EM algorithm indicates that the only pair of likely haplotypes is 3 and 6. In contrast, under the assumption of no LD, all pairs of haplotypes are equally likely. If we code the alleles at the three loci in the pedigree data file as "1 2 1 2 1 2" for both affected brothers, the inferred haplotypes by GENEHUNTER are 1 and 8. However, if we reverse the order of alleles, the inferred haplotypes change. For example, reversing the order for only the first locus (e.g., "2 1") results in haplotypes 4 and 5; reversing the order for the second locus

Table 2

Posterior Probabilities of Haplotype Pairs for a Patient from Example Pedigree

Haplotype Pair	Posterior Probability	Probability if No LD
1 8	.00	.25
3 6	1.00	.25
2 7	.00	.25
4 5	.00	.25

Table 1

Haplotypes and their Frequencies Based on Pedigree Inference and the EM Algorithm Applied to Unrelated Subjects

	Locus Variant			Frequency of Haplotype				
Haplotype	Ile97	Arg462	Glu541	Cases Inferred[a]	Cases EM[b]	Controls EM[b]	Cases, No LD[c]	Controls, No LD[c]
1	1	1	1	.000	.000	.000	.003	.002
2	1	1	2	.000	.000	.000	.003	.003
3	1	2	1	.008	.020	.014	.007	.004
4	1	2	2	.015	.000	.000	.007	.005
5	2	1	1	.118	.000	.000	.143	.157
6	2	1	2	.195	.306	.367	.158	.205
7	2	2	1	.339	.455	.420	.323	.271
8	2	2	2	.324	.219	.199	.357	.353

[a] Haplotypes inferred by GENEHUNTER.

[b] Haplotype frequencies estimated by EM algorithm for case subjects; one affected member of each family was randomly selected for inclusion in the analysis.

[c] Haplotype frequencies for no LD were estimated by the product of allele frequencies.

results in haplotypes 3 and 6; reversing the order of the third locus results in haplotypes 2 and 7. Apparently, when all pairs of haplotypes are considered to be equally likely (because linkage equilibrium is assumed), haplotypes are determined by GENEHUNTER according to the order of alleles in the input file. This feature, combined with the assumption of linkage equilibrium, can grossly mislead investigators who assume the inferred haplotypes are correct. For this example pedigree, we also used SIMWALK2 to infer haplotypes (Sobel and Lange 1996). For our original order of alleles, "1 2 1 2 1 2," the inferred haplotypes were 4 and 5. As with GENEHUNTER, reversing the order of alleles changed the inferred haplotypes. Furthermore, SIMWALK2 uses Markov-chain Monte Carlo and simulated annealing algorithms, which depend on initial values for random seeds; changing these initial seed values can change the inferred haplotypes.

Both GENEHUNTER and SIMWALK2 assume no LD, making it possible to compute either exact (GENEHUNTER) or simulated approximate (SIMWALK2) multipoint linkage statistics, and these programs have proved to be highly effective for these types of analyses. However, users should be cautious about blind use of the inferred haplotypes from this software as if they were directly measured haplotypes. Inferring haplotypes by software that assumes no LD—and then using the inferred haplotypes to measure LD (Abecasis and Cookson 2000)—may be particularly hazardous. Although our example may be extreme, because of small pedigrees with few founders (i.e., little pedigree information from which to infer haplotypes) and because of the large degree of LD among the markers, current evidence points to a frequent occurrence of strong LD among markers spanning short chromosomal regions (Reich et al. 2001).

Rather than inferring the most likely haplotypes under the assumption of no LD and then using these most likely haplotypes as if they were observed, we can establish a better method of analysis by allowing for LD when we estimate haplotype frequencies with pedigree data and then considering all possible haplotype configurations, as was recently implemented for nuclear families (Rohde and Fuerst 2001). The advantage of this approach is that each possible haplotype configuration is weighted according to its likelihood, and likelihood-ratio tests that compare haplotype frequencies between cases and controls would implicitly take into account the increased variance of the haplotype frequency estimates due to linkage-phase ambiguity. However, this approach is computationally intensive for larger pedigrees, because the number of pedigree haplotype configurations is extremely large. For this situation, sensitivity analysis may help to evaluate the reliability of haplotypes inferred from GENEHUNTER or SIMWALK2. If a large number of pedigrees are available, one way to evaluate whether the inferred haplotypes are reasonable would be to compare frequencies of inferred haplotypes with haplotype frequencies based on application of the EM algorithm to pedigree founders or to a random sample of one subject per pedigree, if founders are not available. For large samples, these two procedures should provide similar haplotype frequencies. In addition, one should estimate pairwise LD among the markers, to determine whether any markers are in strong LD, in which case one should be cautious about using software that assumes no LD to infer haplotypes.

Daniel J. Schaid,[1] Shannon K. McDonnell,[1] Liang Wang,[2] Julie M. Cunningham,[2] and Stephen N. Thibodeau[2]

Departments of [1]Health Sciences Research and [2]Laboratory Medicine and Pathology, Mayo Clinic/Foundation, Rochester, MN

References

Abecasis G, Cookson W (2000) GOLD—graphical overview of linkage disequilibrium. Bioinformatics 16:182–183

Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, Faruque M, et al (2002) Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. Nat Genet 30:181–184

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

Fallin D, Schork N (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959

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

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

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

Reich D, Cargill M, Bolk S, Ireland J, Sabei P, Richter D, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander E (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Rohde K, Fuerst R (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Hum Mutat 17:289–295

Schaid D, Rowland C, Tines D, Jacobson R, Poland G (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425–434

Sobel E, Lange K (1996) Descent graphs in pedigree analysis:

applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323–1337

Sobel E, Lange K, O'Connell JR, Weeks DE (1996) Haplotyping algorithms. In: Speed T, Waterman M (eds) Genetic mapping and DNA sequencing: IMA volumes in mathematics and its applications. Springer-Verlag, New York, pp 89–110

Wang L, McDonnell S, Elkins D, Slager S, Christensen E, Marks A, Cunningham J, Peterson B, Jacobsen S, Cerhan J, Blute M, Schaid D, Thibodeau S (2002) Analysis of the HPC1/RNASEL gene in familial and sporadic prostate cancer. Am J Hum Genet 71:116–123

Address for correspondence and reprints: Dr. Daniel J. Schaid, Harwick 775, Section of Biostatistics, Mayo Clinic/Foundation, Rochester, MN 55905. E-mail: schaid@mayo.edu

0002-9297/2002/7104-0030$15.00

Am. J. Hum. Genet. 71:995–996, 2002

Increased Rate of Twins among Affected Sib Pairs

To the Editor:

Recently, Greenberg et al. (2001) and Betancur et al. (2002) reported an excess of twin pairs among affected sib pairs with autism (MIM 209850). Greenberg et al. (2001) reported an excess of both MZ and DZ pairs, whereas Betancur et al. (2002) found an excess of MZ pairs only. Both studies tested the rates of twin pairs among a sample of affected sib pairs against the population rates. The hypothesis put forward was that being a twin is in itself a risk factor for autism. The purpose of this letter is to show that an excess of twin pairs among affected siblings—in particular, an excess of MZ pairs—is what would be expected if genetic factors are implicated in the etiology of a disorder and does not in itself suggest that being a twin confers a risk. Hence, the reported results could be a logical consequence of the affected sibling ascertainment scheme.

The proportion of twin pairs among a random sample of affected siblings from the population depends on the population incidence of twinning and on the concordance rate for the disorder. Let p be the incidence of the disorder in the population; f_{MZ} and f_{DZ} the population rates of MZ twins and DZ twins, respectively; and r_S, r_{DZ}, and r_{MZ} be the (casewise) concordance rates (i.e., the probability that one sibling is affected, given that the other sibling is affected) for nontwin siblings, DZ, and MZ twins, respectively. For each of the three kinds of sib pairs, the probability of 0, 1, and 2 affected individuals is, for $r = r_S$, r_{DZ}, r_{MZ},

$$P(0 \text{ affected}) = (1-p) - p(1-r)$$
$$P(1 \text{ affected}) = 2p(1-r)$$
$$P(2 \text{ affected}) = rp \ .$$

It follows that the proportion of MZ pairs among all pairs of affecteds is

$$f^*_{MZ} = \frac{f_{MZ} r_{MZ}}{f_{MZ} r_{MZ} + f_{DZ} r_{DZ} + (1 - f_{DZ} - f_{MZ}) r_S} \ .$$

Note that this proportion is independent of the population incidence. For small DZ and MZ population rates, $f^*_{MZ} \approx f_{MZ} r_{MZ} / r_S$; that is, we would expect an increase in the rate of MZ twins that is proportional to the increase in the concordance rate relative to nontwin siblings. From epidemiological studies, the estimates for the concordance rates for autism in MZ pairs, DZ pairs, and nontwin siblings are approximately 0.4–0.7, 0.0–0.03, and 0.03, respectively (see Lauritsen and Ewald [2001] and Folstein and Rosen-Sheidley [2001] for reviews), consistent with a very high heritability on a liability scale and the existence of nonadditive genetic variation for liability (see, e.g., Smith 1970). These estimates suggest that the proportion of MZ twin pairs in a random sample of affected sib pairs is approximately 13–23 times larger than the population MZ twinning rate. The observed increases in the MZ rate in the Greenberg et al. (2001) and Betancur et al. (2002) reports are 13 and 16, respectively; they are in accordance with the published concordance rates.

Greenberg et al. (2001) also report a significant increase (a nearly fivefold increase) in the proportion of DZ twins among the affected sib pairs. Estimates of DZ concordance rates have been similar to or lower than the rates among nontwin siblings but have been based on small numbers of observations (Folstein and Rosen-Sheidley 2001; Lauritsen and Ewald 2001). An increase in the rate of DZ twins relative to nontwin siblings could be due to common environmental factors or due to the "stoppage" phenomenon, in which parents with one affected child choose not to have more children. Lastly, Greenberg et al. (2001) compare their observed increased rates of autism in affected twin pairs with the rates for insulin-dependent diabetes mellitus (IDDM). They found a deficit of DZ twin pairs but an excess of MZ twin pairs. These results are also consistent with the genetic epidemiology of IDDM, with reported concordance rates of 0.06, 0.11, and 0.30–0.50 for nontwin siblings, DZ twins, and MZ twins, respectively (see, e.g., Kyvik et al. 1995; Field 2002).