Am. J. Hum. Genet. 62:171–180, 1998

Incorporating Genotypes of Relatives into a Test of Linkage Disequilibrium

Laurent Excoffier[1] and Montgomery Slatkin[2]

[1]Genetics and Biometry Lab, Department of Anthropology and Ecology, University of Geneva, Geneva; and [2]Department of Integrative Biology, University of California, Berkeley

Summary

Genetic data from autosomal loci in diploids generally consist of genotype data for which no phase information is available, making it difficult to implement a test of linkage disequilibrium. In this paper, we describe a test of linkage disequilibrium based on an empirical null distribution of the likelihood of a sample. Information on the genotypes of related individuals is explicitly used to help reconstruct the gametic phase of the independent individuals. Simulation studies show that the present approach improves on estimates of linkage disequilibrium gathered from samples of completely independent individuals but only if some offspring are sampled together with their parents. The failure to incorporate some parents sharply decreases the sensitivity and accuracy of the test. Simulations also show that for multiallelic data (more than two alleles) our testing procedure is not as powerful as an exact test based on known haplotype frequencies, owing to the interaction between departure from Hardy-Weinberg equilibrium and linkage disequilibrium.

Introduction

The extent of linkage disequilibrium between pairs of loci provides useful information about the history of a population, the evolutionary forces governing those loci, and the location of the loci on the chromosomes. In addition to the extent of physical linkage and recombination, admixture, natural selection, mutation, and population growth and decline can all affect the extent of linkage disequilibrium (Weir 1996). Whatever the causes leading to linkage disequilibrium, it is desirable to possess sensitive and powerful methods to detect whether it is present. For instance, accurate and reliable measurements of linkage disequilibrium are definitely needed for important areas of study, such as gene mapping (Chakraborty and Weiss 1988; Hill and Weir 1994; Stephens et al. 1994; Jorde 1995). A test for significant linkage disequilibrium between a pair of loci is a test for nonrandom association of alleles at those loci. Which test is used depends on the data available. If haplotypes are known, either the χ^2 test or Fisher's exact test, which is a better option, can be used (Weir 1996). If only genotypes of unrelated individuals are known, which is often the case for diploid autosomal loci, then the expectation-maximization (EM) algorithm (Dempster et al. 1977) can be applied (Hill 1974, 1975; Long et al. 1995; Slatkin and Excoffier 1996). The problem addressed in this paper is what to do when genotypes of related individuals are included.

Relatives might be included in a sample for two reasons: either they might be part of a sample of genotypes that was obtained for a purpose other than testing for linkage disequilibrium (the data are available and should not be omitted), or an investigator might intentionally include close relatives because their genotypes might help resolve the gametic phase of related individuals but not necessarily all of them. Boehnke (1991) noted that methods for estimation of allele frequencies in pedigrees could be adapted to estimation of haplotype frequencies. However, these estimates of haplotype frequencies cannot then be used as if they were haplotype data, because the uncertainty in those estimates would not be included in the test of linkage disequilibrium. The algorithm described by Boehnke (1991) results in a maximum-likelihood estimate of haplotype frequencies that could be compared with the likelihood of the haplotype frequencies when linkage equilibrium is assumed (that is, haplotype frequencies are simply the product of allele frequencies), to compute a likelihood-ratio test statistic following a χ^2 distribution. Because most studies use sample sizes that are small relative to the number of possible haplotypes, the asymptotic χ^2 approximation for large sample sizes cannot be used in general. In this paper, we use a method that is equivalent to that described by Boehnke (1991), but we implement a parametric bootstrapping procedure to approximate the null distribution of the likelihood-ratio statistic. As has been

Received June 2, 1997; accepted for publication October 22, 1997; electronically published January 2, 1998.

Address for correspondence and reprints: Dr. Laurent Excoffier, Genetics and Biometry Laboratory, Department of Anthropology, CP 511, 1211 Geneva 24, Switzerland. E-mail: Laurent.Excoffier@anthro.unige.ch

0002-9297/98/6201-0025$02.00

Table 1

Relative Probabilities of the Possible Genotype Combinations in a Family of Two Parents and One Offspring

Genotype Combination			Population Probability		Segregation Probability	Relative Probability
Parent 1	Parent 2	Offspring	Parent 1	Parent 2		
A_1B_1/A_2B_2	A_1B_1/A_3B_2	A_1B_1/A_1B_2	$2f_{11}f_{22}$	$2f_{11}f_{32}$	0	0
A_1B_2/A_2B_1	A_1B_1/A_3B_2	A_1B_1/A_1B_2	$2f_{12}f_{21}$	$2f_{11}f_{32}$	1/4	$(f_{12}f_{21}f_{11}f_{32})/(f_{12}f_{21}f_{11}f_{32} + f_{11}f_{22}f_{12}f_{31})$
A_1B_1/A_2B_2	A_1B_2/A_3B_1,	A_1B_1/A_1B_2	$2f_{11}f_{22}$	$2f_{12}f_{31}$	1/4	$(f_{11}f_{22}f_{12}f_{31})/(f_{12}f_{21}f_{11}f_{32} + f_{11}f_{22}f_{12}f_{31})$
A_1B_2/A_2B_1	A_1B_2/A_3B_1	A_1B_1/A_1B_2	$2f_{12}f_{21}$	$2f_{12}f_{31}$	0	0

shown for the case of unrelated individuals (Slatkin and Excoffier 1996), this procedure provides a more reliable test of linkage disequilibrium, from genotype data in samples of related individuals, in the sense that a χ^2 distribution for the likelihood-ratio statistic does not need to be assumed. We then apply this method to simulated data, to compare its performance with that of Fisher's exact test applied to haplotype data and with the EM algorithm applied to genotype data from unrelated individuals. We show that, in general, the inclusion of close relatives in a data set does not provide much additional power, unless both parents and two or more offspring are included.

Methods

Incorporation of Segregation Probabilities into the Likelihood Equation

In this paper, we consider simple cases of samples of related individuals, for which pedigree data are available for only two generations (parents and offspring) and for which the families are completely independent (no half-sibs and no relatedness between the members of the parental generation). The pairs of loci under investigation also are assumed to be so close that no recombination occurs between generations. In this case, a sample is made up of m independent families, and each family is represented by at least two individuals. In a previous study (Slatkin and Excoffier 1996), we considered the special case for which each family is represented by one individual. Genetic information consisted of two-locus genotypes, with the haplotypes for individuals who are doubly heterozygous unknown. The resolution of double heterozygotes into haplotypes is dependent on the unknown two-locus haplotype frequencies ($\boldsymbol{f} = f_1, f_2, \ldots, f_k$), and on the family relationships within the sample. The likelihood of the haplotype frequencies is proportional to

$$L(\boldsymbol{f}) = \Pi_{i=1}^{m} F_i \ , \qquad (1)$$

where F_i is the likelihood for family i. Then, F_i is obtained from

$$F_i = \Sigma_j g_{1j} g_{2j} s_j \ , \qquad (2)$$

where the sum is over all possible genotype combinations of the members of the family and where g_{1j} is the probability of the first parent's genotype, g_{2j} is the probability of the second parent's genotype, and s_j is the probability of Mendelian segregation of the offspring's genotype combination j. Here, the unknown haplotype frequencies are introduced only in the probabilities of the parental genotypes, and the probabilities of the offspring's genotypes depend only on the segregation probabilities of parental haplotypes into the offspring (see table 1).

The number of genotype combinations for each family is the product of the number of possible genotypes for each member of the family. If all independent individuals of the pedigree (the parents) are present in the sample, the total number of genotype combinations will be rather small and will depend mainly on the number of double heterozygotes in the sample. However, if some members of the parental generation are not sampled, then the number of possible genotype combinations may be very large, and this number will depend on the number of alleles (k) at each locus. For instance, consider the case in which $k = 5$ among the sampled individuals. For each locus, we have to allow for the presence of a sixth allele that may be present among the unsampled individuals of the pedigree but that would have escaped detection because it would not have segregated, by chance, in one of the offspring. In this case, there are $6 \times 6 = 36$ potential two-locus haplotypes and $36 \times 37/2 = 666$ potential different genotypes for each unsampled individual. Therefore, in a family in which only full sibs are sampled, $666 \times 667/2 = 222,111$ distinct parental genotype combinations are possible, and the compatibility of all of them must be tested against the offspring's genotypes. In practice, a great many parental genotypes may be eliminated by a prior examination of single-locus incompatibilities between parental and offspring genotypes (e.g., for a description of a simple algorithm, see Lange and Boehnke 1983). Thus, since the complexity of the problem increases with approximately the eighth power of k, we have restricted ourselves to moderately low levels of polymorphism at each locus, with $k \leqslant 5$.

Maximization of the Likelihood, by Use of the EM Algorithm

We used the EM algorithm to find those two-locus haplotype frequencies maximizing the likelihood given by equation (1), as described elsewhere (Excoffier and Slatkin 1995; Slatkin and Excoffier 1996). When k_1 and k_2 alleles at the first and the second locus, respectively, are assumed, the initial expectation (E) step of the algorithm uses randomly assigned haplotype frequencies, represented here by the vector $\boldsymbol{f}^0 = (f_1^0, f_2^0, \ldots, f_K^0)$ of size $K = k_1 k_2$, to compute the expected frequencies of the possible $K(K-1)/2$ parental genotypes $\boldsymbol{g}^0 = (g_{11}^0, g_{12}^0, \ldots, g_{KK}^0)$, where $g_{ij} = 2f_i f_j$ if $i \neq j$ and $g_{ij} = f_i^2$ otherwise.

The maximization (M) step then is performed as a reevaluation of the haplotype frequencies from the relative probabilities of the combinations of parental and offspring genotypes. For example, let us consider a simple family with two parents and one offspring, in which the two-locus phenotypes of the two parents are $A_1A_2B_1B_2$ and $A_1A_3B_1B_2$ and that of the offspring is $A_1A_1B_1B_2$. The different genotype combinations are listed in table 1, with their relative probabilities, for which both the parental population probabilities and the offspring segregation probabilities are involved, following equation (2). These genotype relative probabilities are used as weights associated to the haplotypes involved in the genotypes. A simple counting procedure then explores all genotype combinations and adds up the weights of the haplotypes present in the parents, to produce new sets of haplotype frequencies, $\boldsymbol{f}^1$, which are used in a new E step. The E and M steps are repeated until convergence of haplotype frequencies is reached. In practice, because the EM algorithm does not necessarily lead to the global optimum solution but merely to a local maximum (Excoffier and Slatkin 1995; Long et al. 1995), the EM algorithm is performed several times, with each repetition starting from different initial haplotype frequencies.

Testing for Linkage Disequilibrium, by Use of Pedigree Information

As initially discussed by Hill (1974), if L^* denotes the likelihood derived under the hypothesis of linkage equilibrium, where haplotype frequencies are the products of allele frequencies, $-2\log(L^*/L)$ should asymptotically follow a χ^2 distribution with $(k_1 - 1)(k_2 - 1)$ df, under the hypothesis of linkage equilibrium, where k_1 and k_2 are the number of alleles present at the two loci. However, this asymptotic behavior is only valid for a small number of alleles, and the use of a χ^2 test can lead to a large number of false-significant results in situations of practical interest (Long et al. 1995; Slatkin and Excoffier 1996). Therefore, in practice, it seems necessary to generate the null distribution of L and to use the tail probability $P(\text{random } L \geqslant L)$ as the test probability. Approximate null distributions can be generated for samples of unrelated individuals by random permutation of alleles, at one locus, between individuals (Slatkin and Excoffier 1996). This procedure is not applicable to samples of related individuals, because random permutations of alleles between individuals can lead to incompatible segregation patterns between parents and offspring.

In order to overcome this difficulty, we propose a new procedure using a parametric bootstrap allocation of haplotypes to independent individuals, to generate new random samples of related individuals, under the linkage-equilibrium hypothesis. In more detail, the random samples are generated as follows: The maximum-likelihood allele frequencies at each locus are first obtained by use of the procedure described above, with the individuals' relatedness taken into account, as described in the study by Boehnke (1991). When all parental genotypes are sampled, this step can be replaced by a mere gene-counting procedure, because the offspring's genotypes are not necessary for computation of the allele frequencies. However, this procedure is necessary when one or more independent individuals have been omitted by the sampling process or when there are recessive alleles in the sample. A pool of linkage-equilibrium haplotype frequencies is generated as the product of the allele frequencies. A pair of haplotypes is randomly assigned, by bootstrap, to each independent individual, sampled or not. The independent individuals then are randomly mated to produce their observed offspring. Then, the unobserved independent individuals are finally excluded, in order to produce a sample having exactly the same properties as the original, in terms of size and family relationships. Finally, the likelihood ratio of this sample is evaluated by use of the EM algorithm, as described above. Many samples then can be generated by use of this bootstrap procedure, to approximate the null distribution of the likelihood ratio.

Assessment of the Effect of Different Family Types on the Estimation of Linkage Disequilibrium

In order to assess the effect, on the linkage-disequilibrium test, of different amounts of additional information from relatives in the sample, we studied several simulated samples of 50 diploid individuals, for which we previously had found a discrepancy between the results of an exact test of linkage disequilibrium (described in Slatkin 1994) and those of the test based on the conventional EM algorithm (Slatkin and Excoffier 1996). We arbitrarily selected two cases for which we had significant results with the exact test but nonsignificant results with the EM-based test (i.e., SN results) or for

Table 2

***P* Values for the Linkage-Disequilibrium Test for the Case of $k = 2$**

Case[a]	*P* Value for Exact Test	*P* Value (CV), by Family Type[b]											
		2 Parents and No. of Offspring of					1 Parent and No. of Offspring of				0 Parents and No. of Offspring of		
		0	1	2	3	4	1	2	3	4	2	3	4
NS:													
A	.167	.011	.077	.105	.112	.114	.156	.177	.131	.136	.039	.006	.000
			(.914)	(.576)	(.318)	(.243)	(1.151)	(.998)	(.738)	(.548)	(1.585)	(3.969)	(7.035)
B	.538	.028	.332	.433	.418	.435	.566	.302	.732	.711	.336	.221	.157
			(.656)	(.440)	(.375)	(.234)	(.548)	(.462)	(.300)	(.275)	(.632)	(.631)	(.632)
SN:													
C	.023	.190	.069	.038	.033	.025	.179	.105	.075	.041	.188	.093	.064
			(.656)	(.802)	(.490)	(.412)	(.880)	(1.216)	(1.065)	(.846)	(.856)	(.755)	(.806)
D	.008	.175	.020	.012	.008	.006	.180	.099	.118	.072	.021	.001	.000
			(1.073)	(.848)	(.543)	(.628)	(.999)	(.978)	(1.108)	(.886)	(2.264)	(3.968)	(7.035)

NOTE.—Values that gave a significance result (5% level) opposite to that of the exact test are underlined.

[a] NS = results not significant for the exact test but significant for the EM-based test on unrelated individuals; and SN = results significant for the exact test but not significant for the EM-based test on unrelated individuals. Cases A and B were chosen at random from among the NS cases of a previous study (Slatkin and Excoffier 1996); and cases C and D were chosen at random from among the SN cases of the same previous study.

[b] CV = coefficient of variation.

which we had the opposite results (i.e., NS results), with both $k = 2$ and $k = 5$ alleles per locus. These results were compared with the results obtained for 11 types of samples of related individuals. Each type of sample comprised 25 families in which the parental generation consisted of the same 50 individuals used in the previous study (Slatkin and Excoffier 1996), but the composition of the sampled members of the families differed among the sample types. The 11 family types were as follows: families with two sampled parents, plus either one, two, three, or four offspring; families with only one sampled parent, plus either one, two, three, or four offspring; and families with zero sampled parents but with two, three, or four full siblings. Since the haplotypes present in the parents and used to produce the offspring were identical to those used for the exact test and the conventional EM algorithm, the amount of independent information was kept constant over all analyses. However, the sample size, as measured by the number of sampled individuals, could have varied, of course, in accordance with family size and composition.

For each family configuration, the potential resolution of the parental gametic phase depended on the offspring genotypes. For this reason, we generated for each configuration 100 different reference samples, keeping the same 50 parental genotypes but producing different sets of offspring, by random segregation of parental haplotypes. A null distribution of the likelihood ratio then was obtained empirically for each of these 100 samples of related individuals. Each null distribution was based on either 100 or 1,000 bootstraps (depending on the total required computing time). Thus, the final EM *P* values (shown in tables 2 and 3 and in figs. 1–4) are the *P* values averaged over these 100 null distributions. Comparisons between average *P* values obtained from null distributions based on 100 or 1,000 replicates did not reveal appreciable differences from the mean (results not shown), but the variance was reduced when a larger number of bootstrap replicates were used.

In order to have a more quantitative assessment of the utility of our approach and of its behavior relative to the exact and conventional EM tests, we applied this double-resampling scheme to a larger number of cases. Thus, we randomly selected 100 samples based on 50 independent individuals, for the $k = 2$ and $k = 5$ cases from our earlier study (Slatkin and Excoffier 1996). Only five family configurations (two parents and one offspring, two parents and four offspring, one parent and one offspring, one parent and four offspring, and zero parents and three offspring) were studied for these samples, because of the prohibitive computing time required to analyze all 11 configurations described above.

Results

Estimation of Linkage Disequilibrium for Samples of Different Family Types, in a Few Test Cases

In table 2 we report the results of the tests of linkage disequilibrium for samples of different types of families,

Table 3

***P* Values for the Linkage-Disequilibrium Test for the Case of $k = 5$**

Case	*P* Value for Exact Test	*P* Value (CV), by Family Type											
		2 Parents and No. of Offspring of					1 Parent and No. of Offspring of				0 Parents and No. of Offspring of		
		0	1	2	3	4	1	2	3	4	2	3	4
NS:													
A	.101	.036	.031	.030	.029	.029	.146	.099	.088	.092	.204	.096	.039
			(.013)	(.012)	(.009)	(.007)	(.889)	(.769)	(.583)	(.503)	(.832)	(.894)	(1.300)
B	.283	.003	.144	.129	.134	.126	.257	.218	.307	.338	.241	.175	.098
			(.196)	(.173)	(.162)	(.090)	(.784)	(.708)	(.418)	(.311)	(.848)	(.920)	(.935)
SN:													
C	.0008	.342	.003	.001	.001	.001	.052	.034	.090	.056	.080	.019	.008
			(.777)	(1.013)	(.901)	(.960)	(1.574)	(.774)	(.353)	(.479)	(1.116)	(1.844)	(1.933)
D	.001	.100	.001	.001	.001	.001	.035	.012	.005	.005	.046	.016	.006
			(1.079)	(1.163)	(1.025)	(.997)	(1.598)	(1.709)	(1.833)	(1.957)	(1.369)	(2.399)	(2.208)

NOTE.—See footnotes to table 2.

for $k = 2$. In general, the use of information on relatives improves the results obtained from samples of unrelated individuals, when the results from the exact test are used as the standard of comparison. Improvement is greater when both parents are included in the sample. In this case, the accuracy of the test improved mostly by addition of one offspring, in the sense that the likelihood-ratio test led to conclusions similar to those from the exact test. Addition of a second offspring then led to *P* values that were closer to those of the exact test, whereas addition of additional offspring did not lead to as much improvement. There were, however, a steady increase in the accuracy of the test probability, when the number of offspring was increased, and a reduction in the coefficient of variation. This trend is not as clear in samples comprising families with only one sampled parent or in samples comprising full sibs only. When as many as three offspring were added to the single-sampled-parent families, cases C and D did not reach significance. Addition of a fourth offspring restored significance for case C but not for case D. In families comprising full sibs only, consideration of four offspring did not seem to be enough to lead to *P* values close to those of the exact test. The *P* values also had a much larger variance for the sample configurations with zero parents.

In table 3 we report the results obtained for $k = 5$. Results were also globally better than those of the conventional EM-based test, except for those for case A when two parents were sampled and those for case C when one parent was sampled. Case A led to especially odd results, since one would have expected results closer to those of the exact test when two parents and some offspring are included in the sample than when one or two parents are missing (for a possible explanation of this apparent discrepancy, see the discussion of fig. 4 below). The *P* values obtained with two sampled parents per family always had a smaller variance than those obtained when one or both parents were missing from the sample. This is accounted for by the fact that many parental genotype combinations may be compatible with the offspring's genotypes, when one or two parents are not sampled.

From the small number of cases presented in tables 2 and 3, it appears difficult to recommend a definitive sampling strategy that would optimize both laboratory efforts and testing accuracy. However, compared with a situation in which only independent individuals are sampled, sampling of additional offspring does lead to improved results, and the more offspring that are sampled, the more accurate the results, for a fixed number of sampled parents. However, gathering of samples of full sibs only does not seem to be a good overall strategy, because of the difficulty of correctly inferring the gametic phase of the unsampled parents.

Detailed Comparison of the EM Approach and the Exact Test

In figure 1 we show 100 linkage-disequilibrium *P* values obtained from the EM-algorithm approach, taking into account family relationships, plotted against exact-test *P* values for $k = 2$. For comparison purposes, the *P* values obtained by the EM algorithm, from 50 independent individuals without relatives, are also reported for each case and are shown as unblackened circles. Samples of two parents and one offspring led to *P* values that were in much closer agreement than those of the unrelated-individual case. The correlation for the results of the exact test is stronger (the square of the correlation coefficient $[R^2] = .88$) than that for the conventional

EM-based test of 50 unrelated individuals ($R^2 = .48$). Samples of families with two parents and four offspring led to results virtually identical to those of the exact test ($R^2 = .99$), with the points nicely aligned along the diagonal. Samples of families with one parent and one offspring led to results better than those for samples of independent individuals, in the sense that there was a closer linear relationship between the EM *P* values and the exact-test *P* values ($R^2 = .76$). The overall correlation increases when families of one parent and four offspring are considered ($R^2 = .94$). In contrast to results for samples in which some parents were included, results for families with three full sibs were worse than those for the independent-individual case ($R^2 = .10$ vs. $R^2 = .48$), with no clear relationship between the EM *P* values and the exact-test *P* values.

In figure 2 we show the results obtained for $k = 5$, for family relationships equivalent to those used to obtain the results shown in figure 1. The results are very similar to those for the $k = 2$ case, except with regard to two important points. For samples of families in which both parents were sampled, in the presence of four offspring per family, the EM *P* values and the exact-test *P* values were less correlated than those for the $k = 2$ case, and we observed some false-positive tests, as well as a quite scattered distribution of points around the diagonal ($R^2 = .83$ vs. $R^2 = .99$, for $k = 2$). On the other hand, samples of families comprising three full sibs presented results that improved on those for samples of independent individuals ($R^2 = .74$ vs. $R^2 = .49$), suggesting that, because of the larger number of alleles, parental haplotypes were better resolved than those in the $k = 2$ case.

The slope of the regression of the EM-based *P* values against that of the exact-test *P* values also was different between the two-parent case and the other cases (fig. 2). It clearly approached 1 for the two-parent cases and was less steep for the cases of one sampled parent or zero sampled parents. It had an unexpected effect on the behavior of the test, because it introduced more false-significant results for the two-parent cases than for the one-parent cases, even though the *P* values were globally better estimated for the two-parent cases than for the one-parent cases. This is explained by the overall relative lack of power of the one-parent cases.

Ability of the EM Approach to Recover the Parental Gametic Phase

For the $k = 5$ case, the relative lack of fit between the EM and the exact-test *P* values, when families of two parents and four offspring were sampled, deserved further investigation. For this purpose, we compared the exact-test *P* values with the likelihood-ratio *P* values, obtained by using the information on the gametic phase of the parents, available from our simulated samples. In fact, complete information on the gametic phase of the parents theoretically would be obtained with an infinitely large number of offspring, so that these latter EM *P* values can be seen as the limit that could be obtained by an increase of the family size. Plots of these comparisons are reported in figure 3, for the $k = 2$ and the $k = 5$ cases. Whereas a tight linear relationship ($R^2 = .99$) was found between the exact-test and the likelihood-ratio *P* values for $k = 2$, the distribution for $k = 5$ was more scattered ($R^2 = .82$). This suggests that the lack of fit shown in figure 2 is not attributable to the inability to resolve the parental phase but, rather, to a fundamental difference between the two testing procedures: the exact test of linkage disequilibrium is only based on haplotype and allele frequencies, whereas the likelihood-ratio test also assumes Hardy-Weinberg equilibrium (HWE). Therefore, the *P* value of a likelihood-ratio test will depend not only on the extent of linkage disequilibrium between the two loci but also on possible departure from HWE at either of the two loci or at the haplotype level. Thus, even though the amount of linkage disequilibrium cannot be recovered perfectly, the fact that the *P* values obtained from samples of families with two parents and four offspring were virtually identical to the *P* values computed with known gametic phase (results not shown) implies that the gametic phase and the haplotype frequencies can be recovered efficiently by use of the current approach, for samples of nuclear families.

The impact of departure from HWE was examined further for the $k = 5$ case, by the performance of exact tests of HWE (Guo and Thompson 1992), at both the locus and the haplotype levels. In figure 4, we have superimposed the results of these tests for the 100 cases considered in figure 3 (*right-hand panel*). The cases showing departure from HWE are indicated by blackened markers. We can see that the cases for which the hypothesis of HWE was rejected at the 5% level are spread over the whole distribution and are not only those cases that are far from the diagonal or that resulted in false positives. Thus, departure from HWE was not responsible alone for the false-significant cases in the lower-right quadrant. This result suggests that the discrepancies between the exact test and the likelihood-ratio test were not solely because of departures from HWE but also resulted from interaction between Hardy-Weinberg disequilibrium and linkage disequilibrium. In figure 4 we also show that case A, for which discrepant results are given in table 3, was precisely a case for which the hypothesis of HWE did not hold at the haplotype level, suggesting that the likelihood-ratio test was found to be significant because of the departure from HWE and the departure from linkage disequilibrium. However, note that HWE was observed

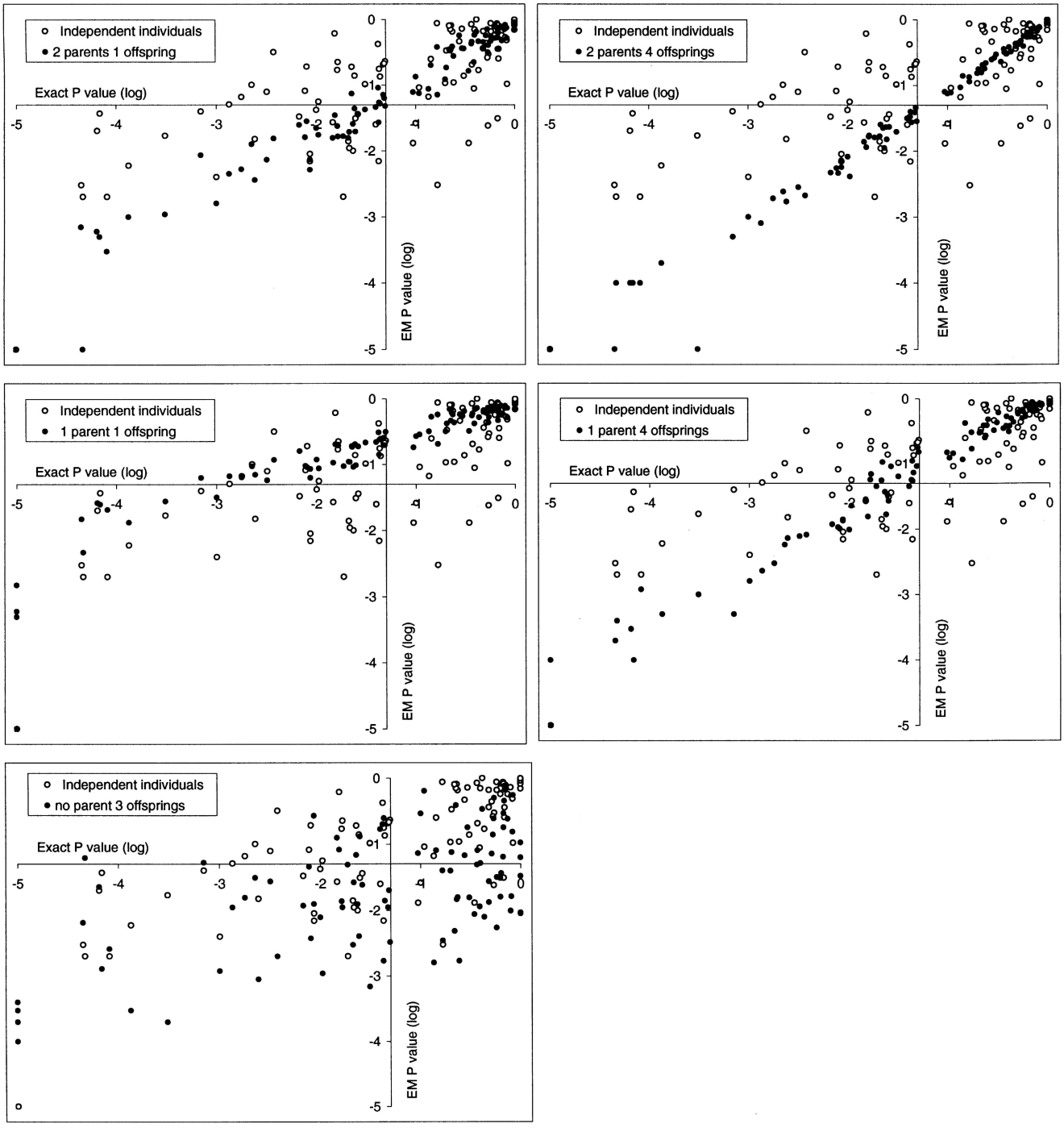

Figure 1 Linkage-disequilibrium *P* values obtained from EM-algorithm approach, taking into account family relationships, plotted against *P* values obtained from the exact test, for $k = 2$ (*blackened circles*). Each point is an average *P* value obtained from 100 null distributions of likelihood ratios obtained from samples of identical parents but offspring generated by random segregation of parental haplotypes, without recombination. Each null distribution was empirically estimated from 100 samples generated under the hypotheses of linkage and HWE (see text). For comparison, EM *P* values computed from the sole independent individuals (parents) are given (*unblackened circles*).

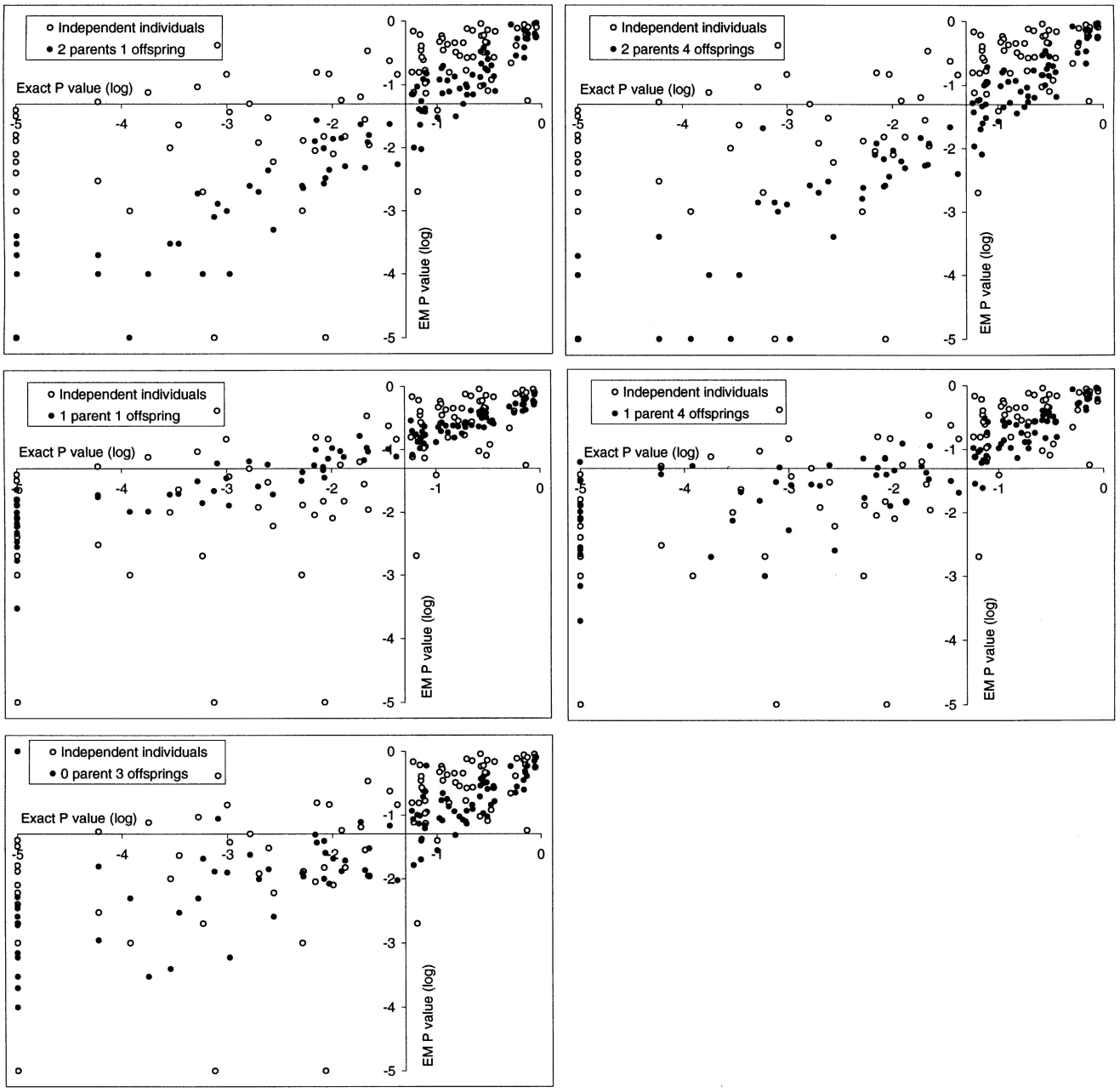

Figure 2 Linkage-disequilibrium *P* values obtained from EM-algorithm approach, taking into account family relationships, plotted against *P* values obtained from the exact test, for $k = 5$. The EM *P* values were obtained in accordance with the procedures described in the legend for figure 1.

at each of the two loci, taken separately, so that, in a practical situation, one would not have been able to attribute the significance to the departure from HWE. If we remove those cases for which HWE was rejected for one reason or another and, thus, consider only those cases shown in figure 4 as unblackend circles, the fit between the exact test and the likelihood-ratio test would improve slightly, increasing from $R^2 = .82$ to $R^2 = .84$.

Discussion

We have shown how to incorporate the genotypes of relatives into the EM algorithm in order to infer hap-

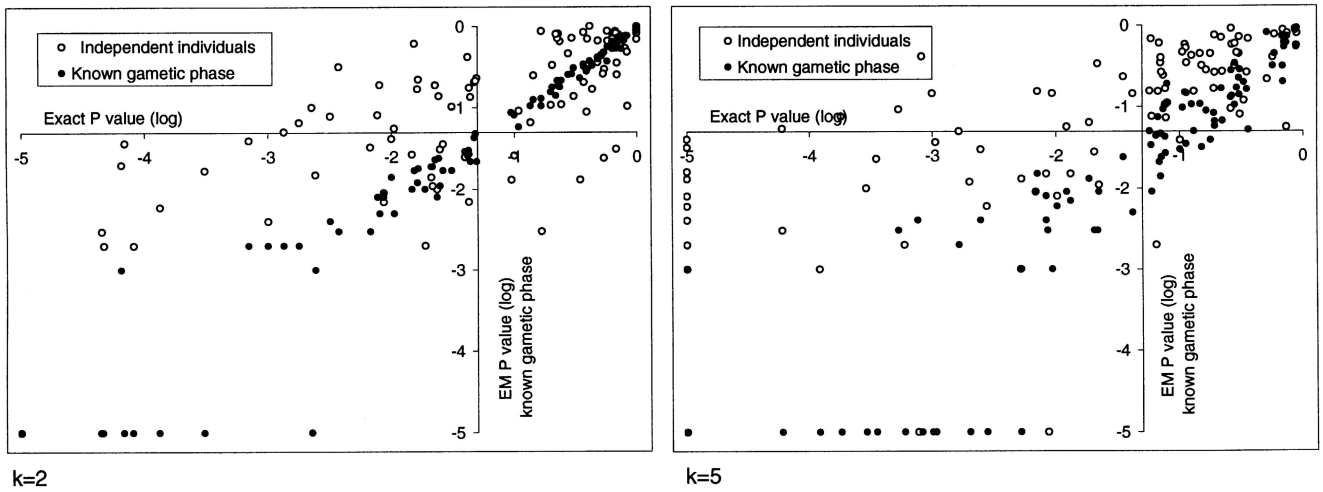

Figure 3 Likelihood-ratio *P* values obtained from the EM-algorithm approach, under the assumption that the gametic phase of the parents is known, plotted against the *P* values obtained from the exact test.

lotype frequencies and to test for significant departures from linkage equilibrium. As expected, the inclusion of close relatives does help resolve the genotypic phases of double heterozygotes, but that effect does not always lead to a substantial increase in the accuracy of the results. When both parents from a family are already represented in a sample, inclusion of some of their offspring always helps. There is less of an improvement when only one parent per family is represented in a sample, and there is no real improvement when only full siblings are sampled. When both parents are not sampled, the uncertainty about the genotypes of the missing parent or parents substantially reduces the parental-phase information that, in principle, is provided by the offspring sampled. Thus, an overall good strategy appears to be to try to accumulate samples of nuclear families comprising independent individuals plus several children. However, families in which only one parent is sampled should not be discarded, since they still improve on the cases in which only unrelated individuals are considered.

For this paper, we investigated simple pedigree structures, but the present method could be extended to accommodate more-complex pedigrees, involving, for instance, one additional generation and half-sibs. The resulting likelihood function to be maximized would be more complex, involving additional genotype combinations per family, but certainly would be manageable, up to a reasonable level. For simplicity, we also assumed no recombination between generations. In principle, recombination fractions can be incorporated into the likelihood function, and likelihood ratios can be computed for different amounts of recombination. In these cases, the resampling procedure would have to allow specifically for recombination when offspring from independent individuals are created. However, the potential interest of population studies that infer genetic linkage is the possibility of using recombination events that have occurred in the ancestry of a sample of genes and not to focus on the last few generations. In this sense, pop-

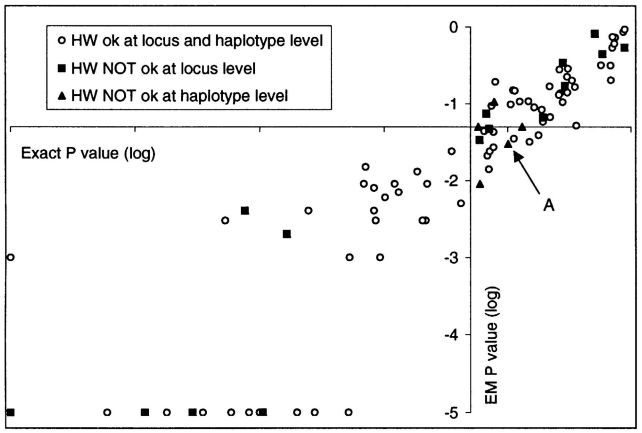

Figure 4 EM *P* values, obtained under the assumption that the parental gametic phase is known, plotted against the exact-test *P* values, for $k = 5$. This plot is partly similar to that shown in the right-hand panel of figure 3, except that those cases for which the hypothesis of HWE is rejected (at level .05) are shown as blackened markers. Unblackened circles indicate cases for which the hypothesis of HWE is accepted, whereas blackened squares indicate those cases for which HWE is rejected for one or both loci separately, and blackened triangles indicate those cases for which HWE is rejected at the haplotype level. The position of case A from table 3 is shown by an arrow.

ulation studies would be most helpful for the detection of linkage disequilibrium between tightly linked markers, when practically no recombination would be observed in any generation. Our current resampling scheme assumes very tight linkage and, thus, is best applied to pairs of loci that are thought to be closely linked.

Although, in this study, we concentrated on cases for which only a relatively low number of alleles were present, our approach can easily handle markers with $k > 5$. In this study, the limit of $k \leqslant 5$ was imposed by the computation burden, since we needed to compute 10,000 null distributions for each type of family indicated in figures 1 and 2. This proved to be particularly computer intensive for the cases including only full sibs, despite the use of the parental genotypic exclusion algorithm provided by Lange and Boehnke (1983). However, when only a few null distributions need to be computed, markers with a larger number of alleles could be accommodated, especially if parents are included in the sample.

One of the implications of our study is the recognition of the importance of a slight departure from HWE in tests of linkage disequilibrium based on genotype data, leading to the rejection of our likelihood-ratio test even in the absence of linkage disequilibrium (as for case A in table 3 and fig. 4). This effect exists irrespective of whether related individuals are incorporated into the sample and appears to be more pronounced with an increase of the number of alleles per locus. This suggests that departure from HWE can be detected more easily with a larger k (as was the case for linkage disequilibrium based on haplotype data [see Slatkin 1994]). It follows that the linkage-disequilibrium test based on genotype data does not asymptotically tend to an exact test, contrary to observations by Hill (1974) for $k = 2$, unless there is perfect random association of gametes. However, even if the association between the exact test and the current procedure is not perfect and false significance can be observed (figs. 1 and 2), an extremely low P value is unlikely to occur in the absence of linkage, for samples of nuclear families (fig. 2). In order to overcome the interference of the HWE assumption, one might be tempted to use only those individuals of the sample whose phase could be unambiguously inferred from the pedigree data. However, it appears difficult to recommend this procedure as a rule, since it would introduce a bias in the estimation both of haplotype frequencies and of linkage disequilibrium, but the extent of the bias remains to be quantified.

Acknowledgments

We thank André Langaney and the anonymous reviewers for their comments and suggestions. This research was supported, in part, by Swiss National Science Foundation grant 32-47053-96 (to L.E.) and U.S. National Institutes of Health grant GM 28428 (to M.S.). A computer program for performance of the computations presented here will be made available from L.E., by request.

References

Boehnke M (1991) Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25

Chakraborty R, Weiss K (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85: 9119–9123

Dempster A, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. J R Stat Soc 39:1–38

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

Guo S, Thompson E (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239

——— (1975) Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics 31:881–888

Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet 54:705–714

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11–14

Lange K, Boehnke M (1983) Extensions to pedigree analysis. V. Optimal calculation of Mendelian likelihoods. Hum Hered 33:291–301

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

Slatkin M (1994) Linkage disequilibrium in growing and stable populations. Genetics 137:331–336

Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the EM algorithm. Heredity 76:377–383

Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55:809–824

Weir BS (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA