Report

Ignoring Linkage Disequilibrium among Tightly Linked Markers Induces False-Positive Evidence of Linkage for Affected Sib Pair Analysis

Qiqing Huang, Sanjay Shete, and Christopher I. Amos

Department of Epidemiology, University of Texas M. D. Anderson Cancer Center, Houston

Most multipoint linkage programs assume linkage equilibrium among the markers being studied. The assumption is appropriate for the study of sparsely spaced markers with intermarker distances exceeding a few centimorgans, because linkage equilibrium is expected over these intervals for almost all populations. However, with recent advancements in high-throughput genotyping technology, much denser markers are available, and linkage disequilibrium (LD) may exist among the markers. Applying linkage analyses that assume linkage equilibrium to dense markers may lead to bias. Here, we demonstrated that, when some or all of the parental genotypes are missing, assuming linkage equilibrium among tightly linked markers where strong LD exists can cause apparent oversharing of multipoint identity by descent (IBD) between sib pairs and false-positive evidence for multipoint model-free linkage analysis of affected sib pair data. LD can also mimic linkage between a disease locus and multiple tightly linked markers, thus causing false-positive evidence of linkage using parametric models, particularly when heterogeneity LOD score approaches are applied. Bias can be eliminated by inclusion of parental genotype data and can be reduced when additional unaffected siblings are included in the analysis.

In multipoint linkage analysis, when there is unresolved phase information for multiple heterozygous individuals, equal probabilities are usually assigned to all possible phases that are compatible with the data (O'Connell and Weeks 1995; Kruglyak et al. 1996). When the markers are sparsely spaced and there is approximate linkage equilibrium, assuming equal phase probabilities does not lead to asymptotic bias. However, this assumption could be problematic when there is strong linkage disequilibrium (LD) among tightly linked markers, because the observed haplotype frequencies deviate from the expected frequencies. Currently, most commonly used linkage programs assume linkage equilibrium between markers and assign equal probabilities to all possible inheritance vectors that explain the data. Applying such programs to markers that are in strong LD can lead to incorrect pedigree haplotype inference (Schaid et al. 2002) and may cause bias in pedigree link-age analysis. Previous analytical studies showed that linkage analysis could be robust to misspecification of phase probabilities (Ott 1999). However, this previous analytical work implicitly assumes both parents are genotyped, and this assumption is often not met. In this report, we demonstrate that assuming linkage equilibrium among markers in LD can induce false-positive evidence for multipoint linkage analysis when one or both parental genotypes are missing.

It has been shown in several studies (Freimer et al. 1993; Williamson and Amos 1995; Knapp et al. 1993) that misspecification of single-marker allele frequencies can lead to false-positive evidence for linkage. In the case of tightly linked loci, haplotypes become analogous to alleles, and, thus, specifying incorrect haplotype probabilities becomes analogous to specifying inaccurate genotype probabilities. However, because of the unknown phases for multiple heterozygous individuals, misspecification of haplotype frequencies is a more complex issue than misspecification of single-allele frequencies. Since inaccurate genotype frequencies cause false-positive evidence for linkage, we decided to study the impact that LD among tightly linked markers may have on linkage analysis, under the usual assumption of linkage equilibrium, which can lead to both misspecification of hap-

lotype frequencies and phase probabilities. When parental data are available, linkage methods use the observed genotypes rather than specified genotype frequencies. In this situation, the genotype frequencies become irrelevant to the analysis, and false-positive results are neither expected nor observed.

LD between tightly linked markers causes certain haplotypes to be more frequent than expected under linkage equilibrium. The accrual of those haplotypes in families may be interpreted as haplotype sharing among family members. In the case of affected sib pair design for linkage analysis, LD can cause apparent oversharing of multipoint identity by descent (IBD) among affected sibs and thus results in false-positive evidence for linkage. As an example, assume that we study two tightly linked markers, each with two alleles, 1 and 2. For these two markers, there are four possible haplotypes: 11, 12, 21, and 22. If these two markers are in complete LD, we can only observe two haplotypes, 11 and 22, and, accordingly, three diplotypes: 11/11, 11/22, and 22/22. If we denote the three diplotypes as 1, 2, and 3, respectively, there are only six possible sib pairs: (1, 1), (1, 2), (1, 3), (2, 2), (2, 3), and (3, 3). In the appendix, we show, in a general way, how to calculate the expected frequencies of each sib pair in terms of haplotype frequencies and calculated multipoint IBD sharing for a sib pair (2, 2), given specific haplotype frequencies. Here we assume equal frequencies of 0.5 each for the two alleles of the two markers, which gives a frequency of 0.5 each for the two possible haplotypes under the assumption of complete LD and a frequency of 0.25 each for the four possible haplotypes under the assumption of no LD. By simply plugging in these numbers, we can calculate the expected IBD sharing for the haplotypes, both under the assumption of complete LD and under that of no LD (table 1). In this example, if LD is taken into account (i.e., if correct haplotype frequencies are provided), the expected proportion of IBD sharing is 0.5. However, if we assume linkage equilibrium, the expected IBD shar-

ing is 0.573. Therefore, we can see that, even if the markers are not linked to the disease locus, LD among markers will cause overestimation of multipoint IBD sharing among sib pairs if linkage equilibrium is assumed. And this bias will generate false-positive evidence of linkage for affected sib pair analysis. Similarly, more than two markers in strong LD will generate even more bias, since the haplotype frequencies will deviate even further from those expected under linkage equilibrium. The magnitude of bias depends on the strength of LD among the markers. To further study the effect of LD on multipoint linkage analysis, we used simulations to study the behavior of IBD-sharing–based, model-free methods of linkage detection by using affected sib pairs with either zero, one, or two parents available for genotyping. We also used parametric linkage analysis as a comparison.

Currently, there are limited programs available to simulate LD between markers within pedigrees. We developed a two-step strategy to simulate LD and pedigree data. First, we simulated LD by randomly assigning haplotypes for all the founders and married-ins, on the basis of specified population haplotype frequencies, which determine the LD. Then we used SLINK (Ott 1989) to simulate segregation of multiple markers conditional on the marker genotypes and disease phenotypes within the pedigrees. We found that this strategy always gives us the desired LD between markers as well as possible recombinants between markers and between marker and disease locus. This approach has no limitation on the number of haplotypes or markers and can be applied to any pedigree structure. The programs and approaches we used are available for download at our Web site.

For this study, we simulated a recessive disease with disease allele frequencies of 1% and 5%, as well as a dominant disease with a disease allele frequency of 1%. We simulated two markers with either equal or unequal allele frequencies under the null hypothesis of no linkage to the disease locus. In the simulation settings, recom-

Table 1

Probability of Multipoint IBD Sharing Between Sib Pairs, Under the Assumptions of Complete LD and No LD

		IBD Sharing, Under the Assumption of					
		LD[b]			No LD[c]		
Sib Pair	Probability[a]	P(IBD = 0)	P(IBD = 1)	P(IBD = 2)	P(IBD = 0)	P(IBD = 1)	P(IBD = 2)
(1, 1)	9/64	1/9	4/9	4/9	1/25	8/25	16/25
(1, 2)	6/32	1/3	2/3	.0	1/3	2/3	.0
(1, 3)	2/64	1	.0	.0	1	.0	.0
(2, 2)	5/16	1/5	2/5	2/5	1/7	2/7	4/7
(2, 3)	6/32	1/3	2/3	.0	1/3	2/3	.0
(3, 3)	9/64	1/9	4/9	4/9	1/25	8/25	16/25

[a] Expected probability when the two markers are in complete LD.
[b] Expected proportion of IBD sharing under the assumption of LD = .5.
[c] Expected proportion of IBD sharing under the assumption of no LD = .573.

bination between the two markers was set to minimal (recombination fraction [θ] 0.001). Allelic association between the two markers varied from no LD to complete LD, as determined by the population haplotype frequencies specified in table 2. We simulated nuclear families, each with one affected sib pair and with either (1) data for neither parent, (2) data for one parent, or (3) data for both parents. For each replicate sample, 1,000 families of the same type were simulated, and 100 replicates were generated for each simulation setting. Both parametric and model-free multipoint and single-point linkage analyses were performed for each data set. Parameters used for linkage analysis (allele frequencies, disease model, penetrance, et al.) were the same as the parameters used for simulations. We used several commonly used linkage programs—Allegro (Gudbjartsson et al. 2000), Merlin (Abecasis et al. 2002), and Genehunter (Kruglyak et al. 1996)—to compare the results. We achieved similar results from different linkage programs and observed similar patterns for different disease models and marker allele frequencies. Therefore, we only present the results obtained from applying Allegro to the analysis of a recessive disease with a disease-causing allele frequency of 1% and equal marker allele frequencies.

Figure 1 depicts the effect of LD on multipoint linkage analysis as a relationship between LD (D') and LOD score (model-free analysis, left panel) or heterogeneity LOD (HLOD) score (parametric analysis, right panels). LOD scores for model-free analysis were obtained using a Kong and Cox exponential model (Kong and Cox 1997) and the score function of S_{pairs} (Whittemore and Halpern 1994). Since these two markers are tightly linked, the LOD scores for these two markers and the intermarker locations were nearly identical, and the maximum LOD score from each replicate was used to represent the evidence of linkage. The average of maximum LOD scores over 100 replicates for each simulation setting was plotted in figure 1. Results were separated for different data sets of affected sib pairs, with either zero (fig. 1*A*), one (fig. 1*B*), or two (fig. 1*C*) pa-

rental genotypes or with one additional unaffected sib with genotypes (fig. 1*D*). When there was no LD ($D' = 0$) between the markers, there was no evidence of linkage. This result showed that the simulation was valid, because the markers were simulated under the null model of no linkage. When neither or only one parent is available for genotyping, LD between markers can cause apparent oversharing of multipoint IBD and positive LOD scores for model-free analysis (fig. 1*A* and 1*B*, left panels). The false-positive evidence for linkage became increasingly extreme as the D' value increased beyond 0.6, which is a value expected for distances of ~100 kb or less (Abecasis et al. 2001; Reich et al. 2001). Whether the false-positive evidence reaches a significant level depends on factors such as the magnitude of LD, allele frequencies, sample size, etc. For example, in our simulations of 1,000 nuclear families, each including one affected sib pair and no available parents, the false-positive rate is ~100% for $D' = 1$, ~90% for $D' = 0.8$, and ~10% for $D' = 0.6$, for a significant LOD score of 3.

When parametric linkage analysis was applied to the simulated data, we found no evidence of linkage (multipoint LOD < -2, for both marker location and intermarker locations). For complex diseases, researchers typically apply HLOD score approaches (Ott 1983; Hodge et al. 2002). Application of heterogeneity linkage analysis to the data resulted in highly positive HLOD scores (fig. 1*A* and 1*B*, right panel). This is not surprising, although analytical calculation of the expected LOD score is excessively complex. Again, let us consider the above example of complete LD between the two markers. The sib pairs (1, 1), (2, 2), and (3, 3) support linkage and provide stronger evidence in favor of linkage under linkage equilibrium than under LD. This is because, in summing over all possible haplotypes in the parents (who have missing genotypes), there are more possible informative genotypes under linkage equilibrium than under the correct LD assumption. The other pairs provide negative evidence for linkage, but the evidence pro-

Table 2

Haplotype Frequencies and LD (D') Settings Between Two Biallelic Markers for the Simulations.

ALLELE FREQUENCIES AND HAPLOTYPE	HAPLOTYPE FREQUENCIES						
	$D' = .0$	$D' = .2$	$D' = .4$	$D' = .6$	$D' = .8$	$D' = .9$	$D' = 1.0$
.5, 0.5:							
11	.250	.300	.350	.400	.450	.475	.500
12	.250	.200	.150	.100	.050	.025	.000
21	.250	.200	.150	.100	.050	.025	.000
22	.250	.300	.350	.400	.450	.475	.500
.2, 0.8:							
11	.040	.072	.104	.136	.168	.184	.200
12	.160	.128	.096	.064	.032	.016	.000
21	.160	.128	.096	.064	.032	.016	.000
22	.640	.672	.704	.736	.768	.784	.800
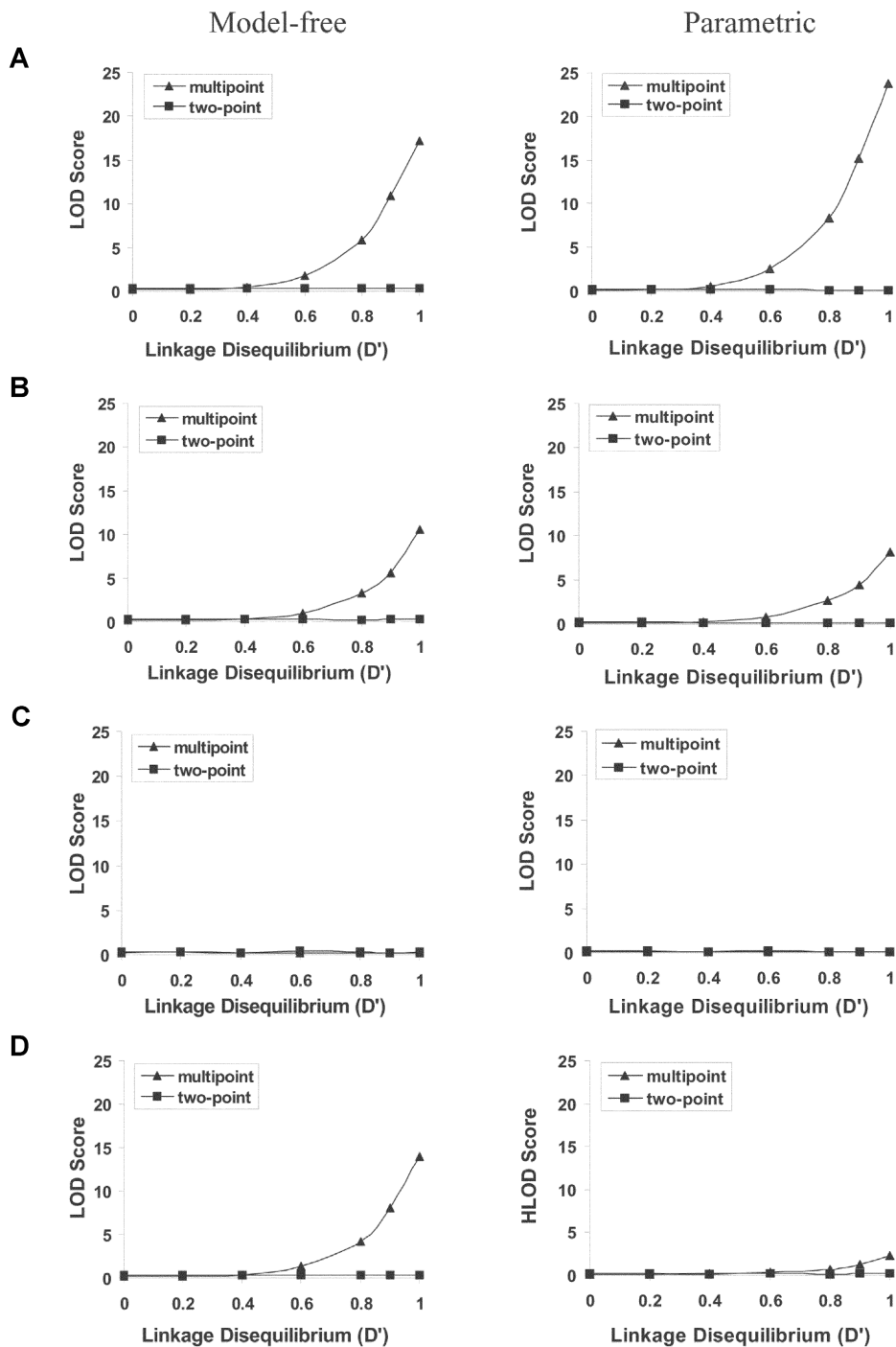
Figure 1 Linkage analyses of two tightly linked markers and an unlinked disease locus. Average maximum LOD scores over 100 replicates were plotted against different magnitudes of LD (D'). Triangles represent the results of multipoint linkage analysis. Squares represent the results of single-point linkage analysis. *Left panels*, LOD scores calculated by model-free linkage analysis. *Right panels*, HLOD scores calculated by parametric linkage analysis. *A*, Affected sib pairs without parental genotypes. *B*, Affected sib pairs with only one parental genotype. *C*, Affected sib pairs with both parental genotypes. *D*, Affected sib pairs with an additional unaffected sib.

vided by the families supporting linkage exceeds that provided by the families negating linkage. The excess information obtained under the erroneous linkage equilibrium assumption leads to false-positive evidence for linkage when an HLOD score is calculated. In addition, similarly high false-positive evidence for linkage is obtained if the maximum LOD score is obtained by varying the distance from the disease locus to the pair of markers (results not presented).

Single-point linkage analysis (parametric and model free) does not suffer from this bias. And bias can be eliminated with parental data (fig. 1*C*). When we added one unaffected sib to the affected sib pair–only data without parental data, the bias for heterogeneity linkage analysis was greatly reduced (fig. 1*D*). However, we still found excess false-positive evidence of linkage for multipoint model-free analysis (fig. 1*D*). The unaffected sibling removed more of the false-positive evidence for linkage in the parametric analysis because we assumed 100% penetrance and no sporadic cases, thus contributing additional negative linkage information. For the model-free analysis, the unaffected sibling(s) only modify the possible genotypes and phase probabilities in the parents, but their phenotype information is not used in the analysis. Adding two unaffected siblings approaches the information provided by having both parents genotyped and thus nearly eliminates false-positive evidence for linkage in the model-free test as well (data not shown). In addition, more markers in strong LD will generate even more bias. For example, when we added one additional marker that is in complete LD with the other two markers, the average LOD score for model-free analysis increased from 17 (panel A in fig. 1) to ~42, and the parametric HLOD score increased from 24 to ~55. In simulation studies for which there was linkage between a disease susceptibility locus and two markers in LD (results not shown), assuming linkage equilibrium increased the LOD scores. However, because the test is no longer valid, these higher LOD scores are not interpretable.

With the advancements in high-throughput genotyping technology, dense markers may be typed in genomic regions without initial evidence of linkage, and multipoint linkage analysis may be performed to detect link-age, with the hope that densely spaced markers (e.g., SNPs) may provide more information than sparse markers (e.g., microsatellites) (John et al. 2004). Our studies indicated that caution should be taken when trying to look for evidence with dense markers where strong LD may exist. The apparent evidence of linkage may reflect an excess of false-positive linkage results due to LD between the tightly linked markers. In the situation of linkage analysis with dense markers, we suggest that evidence of linkage from multipoint linkage analysis should be checked against single-point analysis whenever LD is suspected and that only those markers in low LD should be used for multipoint linkage analysis.

Alternatively, modifications are needed to existing linkage software packages so that they can allow for LD during analysis—for example, by specifying the correct haplotype frequencies. The linkage program LIPED (Ott 1976) can take into account the uncertainty of haplotypes when they are coded as alleles. Therefore, we were able to include LD in multipoint linkage analysis by providing the correct haplotype frequencies. So when we treated each haplotype as an allele and specified the correct haplotype frequencies, bias was eliminated and linkage was excluded for the two tightly linked markers simulated under the null model. But when the haplotype frequencies were specified incorrectly—for example, under the assumption of linkage equilibrium—highly false-positive evidence of linkage occurred, even though the same pedigree data were used. Although LIPED is a good control program to test our hypothesis, it would be tedious to implement for the study of effects from multiple loci. (Interested readers can request the parameter settings for LIPED from the authors.) Current multipoint linkage analysis programs, such as Genehunter, need modification to allow for LD when analyzing multiple SNP markers that are in strong LD.

Acknowledgments

Appendix

To calculate multipoint IBD sharing, all the phase probabilities must be considered. If we know the correct haplotype frequencies, IBD sharing information can be inferred accurately. Unfortunately, current linkage programs assume linkage equilibrium between markers and assign equal probabilities to all possible phases. Here we use a two-marker system as an example to describe how bias can be generated if there is LD between the markers.

For two tightly linked markers, each with two alleles, 1 and 2, there are four possible haplotypes: 11, 12, 21, and 22. The haplotype frequencies are denoted as P_{11}, P_{12}, P_{21}, and P_{22}, respectively. For an individual with genotype 1212, the possible phases are 11/22 and 12/21, with probabilities of $2P_{11}P_{22}$ and $2P_{12}P_{21}$, respectively. For a sib

pair with genotypes of (1212, 1212) and no parental data available, to calculate multipoint IBD sharing for the markers, all possible phase probabilities must be considered. First, we need to calculate the probabilities of the sib-pair data on the basis of $P(G) = P(G|IBD = i)P(IBD = i)$, $i = 0, 1, 2$, where G are the observed genotypes. Let's consider the cases of 0-, 1-, and 2-allele IBD sharing separately.

1. If the two sibs share 0 alleles IBD, there are four possible phase probabilities: (11/22, 11/22), with probability $(2P_{11}P_{22})^2$; (11/22, 12/21), with probability $2P_{11} P_{22} \times 2P_{12} P_{21}$; (12/21, 11/22), with probability $2P_{12} P_{21} \times 2P_{11} P_{22}$; and (12/21, 12/21), with probability $2P_{12}P_{21} \times 2P_{12}P_{21}$.

2. If the two sibs share 1 allele IBD, there are only two phase probabilities: (11/22, 11/22), with probability $P_{11}P_{22} \times P_{22} + P_{11}P_{22} \times P_{11}$, and (12/21, 12/21), with probability $P_{12}P_{21} \times P_{21} + P_{12}P_{21} \times P_{12}$.

3. If the two sibs share 2 alleles IBD, there are also two phase probabilities: (11/22, 11/22), with probability $2P_{11}P_{22}$, and (12/21, 12/21), with probability $2P_{12}P_{21}$.

Obviously, the probabilities depend on the haplotype frequencies. Similarly, we can work out the genotype probabilities for all other sib pairs (among 45 possible sib pairs). For simplicity, we list below only the probabilities for the six possible sib pairs (in no order) under the condition of complete LD between the two markers.

	GENOTYPE PROBABILITY FOR NO. OF ALLELES SHARED IBD		
SIB PAIR	0	1	2
(1111, 1111)	$(P_{11})^2(P_{11})^2$	$(P_{11})^2 \times P_{11}$	$(P_{11})^2$
(1111, 1212)	$2[(P_{11})^2 \times 2P_{11}P_{22} + (P_{11})^2 \times 2P_{12}P_{21}]$	$2[(P_{11})^2 \times P_{22}]$	0
(1111, 2222)	$2(P_{11})^2(P_{22})^2$	0	0
(1212, 1212)	$(2P_{11}P_{22})^2 + 2P_{11}P_{22} \times 2P_{12}P_{21} +$ $2P_{12}P_{21} \times 2P_{11}P_{22} + 2P_{12}P_{21} \times 2P_{12}P_{21}$	$P_{11}P_{22} \times P_{22} + P_{11}P_{22} \times P_{11} +$ $P_{12}P_{21} \times P_{21} + P_{12}P_{21} \times P_{12}$	$2P_{11}P_{22} + 2P_{12}P_{21}$
(1212, 2222)	$2[(P_{22})^2 \times 2P_{11}P_{22} + (P_{22})^2 \times 2P_{12}P_{21})$	$2[(P_{22})^2 \times P_{11}]$	0
(2222, 2222)	$(P_{22})^2(P_{22})^2$	$(P_{22})^2 \times P_{22}$	$(P_{22})^2$

On the basis of the general formula given above, we can plug in the haplotype frequencies and calculate the expected sib-pair probabilities. For example, for sib pair (1212, 1212), we can calculate the probabilities in accordance with different haplotype-frequency settings.

	GENOTYPE PROBABILITY FOR NO. OF ALLELES SHARED IBD			
HAPLOTYPE FREQUENCY	0	1	2	Overall
$P_{11} = P_{22} = P_{21} = P_{22} = .25$ (linkage equilibrium)	1/16	1/16	1/4	7/64
$P_{11} = P_{22} = .5$ (complete LD)	1/4	1/4	1/2	5/16
$P_{11} = P_{22} = .4$, $P_{21} = P_{12} = .1$ ($D' = .6$)	289/2,500	13/100	34/100	1,789/10,000

Then we can calculate the IBD sharing for such a sib pair, using $P(IBD = i|G) = P(G|IBD = i)P(IBD = i)/P(G)$, $i = 0, 1, 2$, and we can create an IBD sharing table:

Haplotype Frequency	P(IBD = 0)	P(IBD = 1)	P(IBD = 2)	π
$P_{11} = P_{22} = P_{21} = P_{22} = .25$ (linkage equilibrium)	1/7	2/7	4/7	.714
$P_{11} = P_{22} = .5$ (complete LD)	1/5	2/5	2/5	.6
$P_{11} = P_{22} = .4$, $P_{21} = P_{12} = .1$ ($D' = .6$)	289/1,789	650/1,789	850/1,789	.657

From this table, we can see a pattern. If the two markers are really in complete LD but we provide the wrong haplotype frequencies in the data analysis, the more bias will be generated the further the specified haplotype frequencies deviate from the true ones. And assuming linkage equilibrium always creates an upward bias. A specific example was given in the text.

Electronic-Database Information

The URL for data presented herein is as follows:

Authors' Web site, http://request.mdacc.tmc.edu/˜qhuang/Software/pub.htm

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO (2001) Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191–197

Freimer NB, Sandkuijl LA, Blower SM (1993) Incorrect specification of marker allele frequencies: effects on linkage analysis. Am J Hum Genet 52:1102–1110

Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12–13

Hodge SE, Vieland VJ, Greenberg DA (2002) HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity. Am J Hum Genet 70:556–559

John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A, Eyre S, Jones KW, Ollier W, Silman A, Gibson N, Worthington J, Kennedy GC (2004) Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. Am J Hum Genet 75:54–64

Knapp M, Seuchter SA, Baur MP (1993) The effect of misspecifying allele frequencies in incompletely typed families. Genet Epidemiol 10:413–418

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188

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

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402–408

Ott J (1976) A computer program for linkage analysis of general human pedigrees. Am J Hum Genet 28:528–529

——— (1983) Linkage analysis and family classification under heterogeneity. Ann Hum Genet 47:311–320

——— (1989) Computer simulation methods in human linkage analysis. Proc Natl Acad Sci USA 86:4175–4178

——— (1999) Analysis of human genetic linkage, 3rd edition. Johns Hopkins University Press, Baltimore, MD, p 251

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN (2002) Caution on pedigree haplotype inference with software that assumes linkage equilibrium. Am J Hum Genet 71:992–995

Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127

Williamson JA, Amos CI (1995) Robustness of the guess LOD approach. Genet Epidemiol 12:163–176