Accurate Inference of Relationships in Sib-Pair Linkage Studies

Michael Boehnke[1] and Nancy J. Cox[2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor; and [2]Department of Medicine, University of Chicago, Chicago

Summary

Relative-pair designs are routinely employed in linkage studies of complex genetic diseases and quantitative traits. Valid application of these methods requires correct specification of the relationships of the pairs. For example, within a sibship, presumed full sibs actually might be MZ twins, half sibs, or unrelated. Misclassification of half-sib pairs or unrelated individuals as full sibs can result in reduced power to detect linkage. When other family members, such as parents or additional siblings, are available, incorrectly specified relationships usually will be detected through apparent incompatibilities with Mendelian inheritance. Without other family members, sibling relationships cannot be determined absolutely, but they still can be inferred probabilistically if sufficient genetic marker data are available. In this paper, we describe a simple likelihood ratio method to infer the true relationship of a putative sibling pair. We explore the number of markers required to accurately infer relationships typically encountered in a sib-pair study, as a function of marker allele frequencies, marker spacing, and genotyping error rate, and we conclude that very accurate inference of relationships can be achieved, given the marker data from even part of a genome scan. We compare our method to related methods of relationship inference that have been suggested. Finally, we demonstrate the value of excluding non–full sibs in a genetic linkage study of non–insulin-dependent diabetes mellitus.

Introduction

Relative-pair methods, such as those based on affected sib pairs (ASPs) (e.g., see Blackwelder and Elston 1985), affected relative pairs (e.g., see Weeks and Lange 1988), and discordant sib pairs (Risch and Zhang 1995), are standard tools for linkage studies of complex genetic diseases and quantitative traits. These methods assess the sharing of marker alleles identical by descent (IBD) or identical by state (IBS) between relative pairs and conclude that there is linkage if observed sharing is sufficiently greater (or less, in the case of discordant sib pairs) than sharing expected under the assumption of no linkage. Since expected sharing depends on the relationship of the pairs, accurate knowledge of these relationships is critical if valid inference is to be achieved.

Given additional genotyped family members, incorrect specification of relationship may be detected on the basis of apparent incompatibilities with Mendelian inheritance. For example, a supposed parent and offspring may fail to share an allele at a marker locus, or three supposed full sibs may between them possess five or six alleles for a single marker. In the absence of additional family members, incorrect specification of sibling relationships cannot be determined with certainty but can be inferred probabilistically on the basis of the frequency with which the pairs share marker alleles.

In this paper we describe a simple likelihood ratio method to infer genetic relationships on the basis of genetic marker data. The method compares the multipoint probability of the marker data, conditional on different genetic relationships, and it infers that relationship that makes the data most likely. This method was previously described by Thompson (1975) for unlinked markers in the more general context of inferring genealogies and has been extended by Goring and Ott (1995) to allow for linkage in a Bayesian framework. Such a method can be employed to exclude probable non–full sibs from an analysis or to establish samples of full-sib pairs and half-sib pairs that can be analyzed separately. This strategy can increase the power to detect linkage in genetic mapping studies.

We discuss the accuracy of our likelihood ratio method to correctly classify the types of relationships that most likely would occur in a putative sibship: MZ twins (or unintentionally duplicated samples), full-sib pairs, half-sib pairs, and unrelated pairs. We address the effects of number of markers, marker spacing, marker allele frequencies, and genotyping error on the analysis. We also note that the likelihood ratio method is, in general, more accurate than related methods based on the observed numbers of marker alleles IBS (Chakraborty and Jin 1993*a*, 1993*b;* Ehm and Wagner 1996; Stivers et al. 1996), although these methods also are quite accurate when many genetic markers are typed.

Received January 15, 1997; accepted for publication May 12, 1997.

Address for correspondence and reprints: Dr. Michael Boehnke, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: boehnke@umich.edu

0002-9297/97/6102-0022$02.00

Methods

Assumptions

Let X_{k1} and X_{k2} be the genotypes at marker k ($1 \leq k \leq M$) for a relative pair, and let $X_k = (X_{k1},X_{k2})$. Assume that the markers are all codominant and autosomal and that the corresponding allele frequencies $q_{k\ell}$ ($1 \leq \ell \leq n_k$) and recombination fractions θ_k ($1 \leq k \leq M-1$) are known without error. Furthermore, let $\psi_k = \theta_k^2 + (1-\theta_k)^2$.

Method Based on the Probability of the Marker Data

To infer the relationship of the pair, we calculate the multipoint probability $P(X|R)$ of the observed marker genotypes $X = (X_1, \ldots, X_M)$, conditional on each relationship R to be considered. We then infer the relationship R^* among those considered for which the probability of the marker data is maximum. For putative full-sib pairs, this might include MZ twins, full sibs, half sibs, and unrelated individuals. Given only two relationships, R_1 and R_2—for example, full sibs and half sibs—the evidence is conveniently summarized as a likelihood ratio, $\mathrm{LR}(R_1,R_2) = P(X|R_1)/P(X|R_2)$, with values $\mathrm{LR} > 1$ suggesting R_1 and $\mathrm{LR} < 1$ suggesting R_2.

To calculate $P(X|R)$, let I_{kf} (I_{km}) be 1 or 0, depending on whether the relative pair shares or fails to share their paternal (maternal) allele at marker k IBD, and let $I_k = (I_{kf},I_{km})$ and $I = (I_1, \ldots, I_M)$. Define $\alpha_k(j|R) = P(X_1, X_2, \ldots, X_{k-1}, I_k = j|R)$ to be the joint probability of the data for the first $k-1$ markers and the IBD-status vector $I_k = j$ at marker k. Recursive calculation of $\alpha_1, \alpha_2, \ldots, \alpha_M$ permits the rapid evaluation of $P(X|R)$ for any noninbred relationship R by making use of the fact that IBD-status vectors $I_1, I_2, \ldots, I_M$ are a hidden Markov chain.

For the first marker, $\alpha_1(j|R) = P(I_1 = j|R)$. For full sibs, $\alpha_1(j|R)$ takes on the values ($^1/_4$,$^1/_4$,$^1/_4$,$^1/_4$) for j = (0,0), (0,1), (1,0), and (1,1), respectively; these probabilities are ($^1/_2$,$^1/_2$,0,0) for (maternal) half sibs, (0,0,0,1) for MZ twins, and (1,0,0,0) for unrelated individuals.

For subsequent markers, the update formula for the recursion is $\alpha_{k+1}(j|R) = \Sigma_i\ \alpha_k(i|R)\ t_k(i,j)\ P(X_k|I_k{=}i)$. Here, $P(X_k|I_k{=}i)$ is the conditional probability of the data at marker k, given the IBD status of the pair; these probabilities are displayed in table 1 (Thompson 1975; Risch 1990). $t_k(i,j) = P(I_{k+1}{=}j|I_k{=}i,R)$ is the probability of moving from IBD-status vector $i = (i_1,i_2)$ at marker k to IBD-status vector $j = (j_1,j_2)$ at marker $k+1$. For full sibs, $t_k(i,j) = (1-\psi_k)^{|j_1-i_1|+|j_2-i_2|}\psi_k^{2-|j_1-i_1|-|j_2-i_2|}$. For maternal half sibs, $t_k(i,j) = (1-\psi_k)^{|j_2-i_2|}\psi_k^{1-|j_2-i_2|}$. For MZ twins, $t_k(i,j) = 1$ for $i = j = (1,1)$. For unrelated individuals, $t_k(i,j) = 1$ for $i = j = (0,0)$.

The final summation $\Sigma_j\ \alpha_M(j|R)\ P(X_M|I_M{=}j)$ yields $P(X|R)$. This sort of recursive strategy to calculate the probability for a hidden Markov chain was first described by Baum (1972) in the context of signal processing. It has been employed to solve a number of problems in genetic analysis (e.g., see Kruglyak and Lander 1995; Lange et al. 1995).

Methods Based on the Number of Marker Alleles IBS

Ehm and Wagner (1996) and Stivers et al. (1996), building on the previous work of Chakraborty and Jin (1993*a,* 1993*b*), recently described methods to infer relationships on the basis of the number of marker alleles IBS in a relative pair. They calculate a sum of the form $S = \Sigma_k\ S_k(X_k)$, where the sum is over all M markers, and $S_k(X_k)$ is a score based on the proportion of marker alleles shared by the genotypes X_{k1} and X_{k2} at marker k; the scores they used are displayed in table 1. Ehm and Wagner and Stivers et al. then calculate a test statistic of the form $Z = [S-\mathrm{E}(S|R)]/\mathrm{SD}(S|R)$, where $\mathrm{E}(S|R)$ and $\mathrm{SD}(S|R)$ are the mean and SD of S, conditional on the relationship R. In sufficiently large samples, Z should be approximately distributed as standard normal if the assumed relationship R is correct. This permits a hypothesis-testing or interval-estimation approach to assess whether a particular relationship, such as full sibs, is correct.

Assessing Methods by Computer Simulation and Application to Non–Insulin-Dependent Diabetes Mellitus (NIDDM)

To assess the accuracy of classification of relative pairs by our method and to compare the accuracy of our method to those of Ehm and Wagner (1996) and Stivers et al. (1996), we performed a computer simulation. For markers equally spaced along the autosomal genome at 10- or 20-cM intervals, we generated 10,000 pairs each of full sibs, half sibs, and unrelated individuals. Markers had either four equally frequent alleles or seven alleles with frequencies .40, .20, .20, .05, .05, .05, and .05, as might be observed for a microsatellite repeat; heterozygosity (H) for each marker type was .75. Markers were placed on each autosome, beginning with chromosome 1 and proceeding through the chromosomes in increasing numerical order; for each chromosome, markers were placed beginning at one telomere and at equal intervals along the chromosome until no more markers could be placed. We used the chromosome-length estimates of Morton (1991) and the corresponding autosomal genome length of 3,854 cM and applied Kosambi's (1944) mapping function to relate map distance and recombination fraction. For each simulation condition, we then calculated the proportion of times that the correct relationship (full sibs, half sibs, unrelated pairs, or MZ twins) was chosen by our likelihood ratio method.

To compare our method with those of Ehm and Wagner (1996) and Stivers et al. (1996), we restricted our attention to full-sib pairs and half-sib pairs. For full-

Table 1

Probabilities and IBS Scores for Genotype Pairs

GENOTYPE[a]		$P(X_1, X_2 \mid I)$, FOR I =			SCORES[b]	
X_1	X_2	(0, 0)	(0, 1), (1, 0)	(1, 1)	S_{EW}	S_{ST}
ii	*ii*	q_i^4	q_i^3	q_i^2	1	1
ii	*ij*	$4q_i^3q_j$	$2q_i^2q_j$	0	½	⅔
ii	*jj*	$2q_i^2q_j^2$	0	0	0	0
ii	*jk*	$4q_i^2q_jq_k$	0	0	0	0
ij	*ij*	$4q_i^2q_j^2$	$q_iq_j(q_i + q_j)$	$2q_iq_j$	1	1
ij	*ik*	$8q_i^2q_jq_k$	$2q_iq_jq_k$	0	½	½
ij	*kℓ*	$8q_iq_jq_kq_\ell$	0	0	0	0

[a] i, j, k, and ℓ are assumed to be distinct alleles at a single genetic marker; X_1 and X_2 are the genotypes for the relative pair at that marker.

[b] IBS scores used by Ehm and Wagner (1996) and Stivers et al. (1996), respectively.

sib–pair data, we determined the critical values for the allele-sharing methods that resulted in approximately the same misclassification rate of full-sib pairs as half-sib pairs as was seen in our method. We then applied these critical values to the test statistics obtained for the half-sib data and compared the resulting rate of misclassification of half-sib pairs as full sibs for each of the three methods.

We also applied our likelihood ratio method to the NIDDM mapping data of Hanis et al. (1996). They reported the results of a genome scan for NIDDM, based on a primary set of 346 Mexican American ASPs from 176 independent sibships; all families were from Starr County, Texas. In their study, the strongest evidence for linkage was found with marker D2S125 on chromosome 2q. As part of their analysis, Hanis et al. (1996) used an IBS-scoring method (see below) to identify and exclude probable non–full sibs.

Results

Accuracy of Classification

Table 2 displays the estimated probability of classifying relative pairs as full sibs, half sibs, or unrelated, for different numbers of markers and marker spacings, assuming equally spaced markers with four equally frequent alleles and no genotyping error. As expected, more markers or greater distances between markers (given a fixed number of markers) resulted in lower probabilities of misclassification. Genotype data from a 20-cM genome scan (206 markers) resulted in misclassification rate estimates of .0006, .0020, and .0008 for full-sib pairs, half-sib pairs, and unrelated pairs, respectively; a 10-cM genome scan (399 markers) reduced these estimates to 0. Even a half-genome scan resulted in low misclassification rates: .0090, .0248, and .0156 for a 20-cM map (100 markers) and .0017, .0030, and .0010 for a 10-cM map (200 markers). Given markers with unequal allele frequencies, our method generally resulted in slightly lower misclassification-probability estimates (data not shown).

Table 3 addresses the impact of genotyping error on relationship-misclassification rates, assuming an allele-typing-error rate of 1%, or a genotype-error rate of essentially 2%. Although misclassification rates were increased over those estimated under the assumption of no genotyping error, the increases were generally modest. For example, for half-genome scans of 100 markers spaced at 20 cM and of 200 markers spaced at 10 cM, misclassification-rate estimates for full-sib pairs increased from .0090 to .0135 and from .0017 to .0024, respectively. With higher genotyping-error rates, the method still can be useful for identifying relationships, although performance does degrade with increasing error rate (data not shown).

Comparison with the IBS-Scoring Methods

A comparison of the results from our method and those of the IBS-based methods of Ehm and Wagner (1996) and Stivers et al. (1996) is presented in table 4, for markers equally spaced at 10-cM intervals. For all combinations of marker type (equal or unequal allele frequencies), number of markers, and genotyping-error rate, our likelihood ratio method resulted in lower misclassification rates than those produced by the IBS-based methods. These differences were largest for unequal marker allele frequencies. The advantage of the likelihood ratio method was greater still for unequally spaced markers (data not shown).

Application to NIDDM

To assess the possible impact of our method on an actual linkage study, we applied it to the Mexican American ASP sample described by Hanis et al. (1996). The

Table 2

Likelihood-Ratio-Classification Probability Estimates: No Genotyping Error

Distance (in cM), no. of Markers, and True *R*	Inferred *R*		
	Full Sibs	Half Sibs	Unrelated
10:			
50:			
Full sibs	.9140	.0850	.0010
Half sibs	.0437	.8722	.0841
Unrelated	.0001	.0585	.9414
100:			
Full sibs	.9809	.0191	.0000
Half sibs	.0100	.9649	.0251
Unrelated	.0000	.0170	.9830
200:			
Full sibs	.9983	.0017	.0000
Half sibs	.0010	.9970	.0020
Unrelated	.0000	.0010	.9990
300:			
Full sibs	1.0000	.0000	.0000
Half sibs	.0001	.9999	.0000
Unrelated	.0000	.0002	.9998
399:			
Full sibs	1.0000	.0000	.0000
Half sibs	.0000	1.0000	.0000
Unrelated	.0000	.0000	1.0000
20:			
20:			
Full sibs	.8384	.1539	.0077
Half sibs	.1161	.7078	.1761
Unrelated	.0047	.1518	.8435
50:			
Full sibs	.9481	.0519	.0000
Half sibs	.0377	.8986	.0637
Unrelated	.0001	.0621	.9378
100:			
Full sibs	.9910	.0090	.0000
Half sibs	.0074	.9752	.0174
Unrelated	.0000	.0156	.9844
150:			
Full sibs	.9983	.0017	.0000
Half sibs	.0009	.9938	.0053
Unrelated	.0000	.0042	.9958
206:			
Full sibs	.9994	.0006	.0000
Half sibs	.0004	.9980	.0016
Unrelated	.0000	.0008	.9992

NOTE.—Markers each have four equally frequent alleles. Estimates are based on 10,000 simulated pairs each.

results that we report differ slightly from those reported by Hanis et al. (1996), who used the related IBS-based method of Chakraborty and Jin (1993*a,* 1993*b*), and reflect the availability to us of additional genotype data, for a total of 455 autosomal markers. Among 346 putative full-sib pairs, our likelihood ratio method and the IBS-based method of Stivers et al. (1996) classified 8 pairs as half sibs, 1 pair as unrelated, and 1 pair as MZ twins (or inadvertent sample duplication). Two of the pairs identified by our method as full sibs were identified as half sibs by the IBS-based method, and two of the pairs identified by the IBS-based method as full sibs were identified as half sibs by our method.

When all 346 ASPs were included within the primary set, the maximum LOD score for D2S125 was 2.96. We then excluded the individuals who appeared not to be full sibs by our likelihood ratio method, resulting in exclusion of seven complete sibships (six of size two and one of size three); additionally, one individual was removed from each of two sibships of size three. Excluding these individuals yielded a maximum LOD score of 3.15 (an increase of 5.1%) and an increase in the estimated IBD sharing, from .61 to .62.

Discussion

In linkage studies of complex genetic diseases and quantitative traits, we generally attempt to localize genes of modest effect, and large numbers of families usually are required. Identifying likely non–full-sib pairs in a sib-pair study is a simple procedure that requires no additional molecular work and only trivial additional statistical analysis; <1 min was required to carry out the NIDDM analysis on a SUN SPARC 10 computer,

Table 3

Likelihood-Ratio-Classification Probability Estimates: 2% Genotyping Error

Distance (in cM), no. of Markers, and True *R*	Inferred *R*		
	Full Sibs	Half Sibs	Unrelated
10:			
100:			
Full sibs	.9703	.0297	.0000
Half sibs	.0078	.9642	.0280
Unrelated	.0000	.0150	.9850
200:			
Full sibs	.9976	.0024	.0000
Half sibs	.0003	.9964	.0033
Unrelated	.0000	.0017	.9983
399:			
Full sibs	1.0000	.0000	.0000
Half sibs	.0000	1.0000	.0000
Unrelated	.0000	.0000	1.0000
20:			
100:			
Full sibs	.9865	.0135	.0000
Half sibs	.0046	.9730	.0224
Unrelated	.0000	.0156	.9844
206:			
Full sibs	.9992	.0008	.0000
Half sibs	.0000	.9979	.0021
Unrelated	.0000	.0014	.9986

NOTE.—See footnote to table 2.

Table 4

Comparison of Methods: Accuracy of Classification of Full Sibs (FS) and Half Sibs (HS)

No. of Markers, Genotype-Error Rate, and Marker-Allele Frequency	P (classify HS as FS by LR)[a]	P (classify FS as HS)[b]		
		by LR	by ST	by EW
100:				
.00:				
Equal	.0191	.0100	.0603	.1117
Unequal	.0150	.0090	.0872	.1317
.02:				
Equal	.0297	.0078	.0507	.0841
Unequal	.0227	.0073	.0678	.1094
200:				
.00:				
Equal	.0017	.0010	.0146	.0364
Unequal	.0005	.0006	.0910	.1208
.02:				
Equal	.0024	.0003	.0241	.0508
Unequal	.0017	.0001	.0235	.0427

Note.—Markers are equally spaced at 10-cM intervals. Estimates are based on 10,000 simulated pairs each.

[a] Fraction of simulated FS misclassified as HS by our likelihood-ratio method (LR).

[b] Fraction of simulated HS misclassified as FS. For the methods of Stivers et al. (ST) and Ehm and Wagner (EW), the critical value for the test was that which resulted in the same misclassification rate of HS as FS as was observed with our likelihood-ratio method (LR).

by either our method or that of Stivers et al. (1996). The advantage of excluding probable non–full-sib pairs was demonstrated by the results from the NIDDM study (Hanis et al. 1996). Alternatively, it may be useful to include half-sib pairs in an analysis that correctly takes into account their relationships; this will be particularly true if the number of probable half-sib pairs is large.

Our simulation results demonstrate that full-sib pairs, half-sib pairs, and unrelated pairs can be accurately differentiated by use of our likelihood ratio method. Eliminating putative sib pairs for which another relationship is more likely should result in only a few true full-sib pairs being excluded when even a portion of a genome scan has been completed. Analysis of 100, 200, and 399 markers in a 10-cM map resulted in an estimated fraction of, respectively, <.03, <.003, and <.0001 full-sib pairs being excluded even when a 2% genotype-error rate was assumed, while eliminating nearly all half-sib pairs and unrelated pairs. Earlier in a genome scan, when fewer markers have been genotyped, eliminating only those pairs for which the data are substantially less likely when a full-sib relationship is assumed than when some other relationship is assumed should still eliminate many non–full-sib pairs, while sacrificing few true full-sib pairs. Indeed, since misclassification of full sibs as half sibs is more common than missclassification of half sibs as full sibs, and since full sibs usually are more common than half sibs, eliminating only those putative full-sib pairs for which the data are substantially more likely when half sibs are assumed might be a good general strategy.

The accuracy of any marker-based method to infer relationships must strongly depend on the degree of polymorphism of the markers. We have concentrated on the case of markers with $H = .75$, typical for the microsatellite markers that are the workhorses for current gene-mapping studies. In the limit of completely informative markers ($H = 1.00$), accurate identification of relatives can be achieved with only a few markers. For example, accurate genotyping of 22 unlinked markers results in a misclassification rate for full-sib pairs as half sibs of $(\frac{3}{4})^{22} - (\frac{1}{2})^{22} < .002$ and a misclassification rate of half-sib pairs as full sibs or unrelated pairs of no more than $(\frac{1}{2})^{22} < .000001$. Given 2% genotyping error, simulation of typing 50 markers at the beginning of a 10-cM genome scan results in a misclassification rate of .0163 for full-sib pairs as half sibs and of .0055 for half sibs as either full sibs or unrelated.

In contrast, given biallelic markers, more markers are required. Still, in simulations with 200 or 400 markers with two equally frequent alleles in a 5-cM scan, only .0218 or .0021, respectively, of full-sib pairs are misclassified as half sibs, and only .0152 or .0017, respectively, of half-sib pairs are misclassified as full sibs. Thus, the probable move toward gene mapping by use of large numbers of inexpensive biallelic markers still will permit accurate inference of relationships, since the large number of markers required for the linkage analysis will, in turn, be sufficient to allow accurate inference of relationships.

In our simulations, we allowed for the possibility of genotyping error but assumed that marker allele frequencies, marker order, and distances between the markers all were known without error. Although these assumptions will not all hold, they all should be well approximated. So long as marker allele frequencies are estimated from the family data (Boehnke 1991), those estimates should be quite accurate, particularly given the large number of sibling pairs generally required for mapping genes for complex traits. For a 10- or 20-cM map of markers typed on the CEPH reference pedigrees or the subset of the largest such pedigrees, marker-ordering errors are rare, and distance estimates generally are quite accurate. Data from densely mapped regions probably should not be included in the identification of non–full sibs, since order will be less certain, since little additional information will be gained because IBD status of relative pairs for tightly linked markers are highly correlated, and since regions that are densely mapped are explored more intensively precisely because of their evidence for linkage.

Although in our simulations we have concentrated on misclassification of full-sib pairs, half-sib pairs, and unrelated pairs, the unknown presence of MZ twins or of duplicated samples also is of potential concern, since it generally will spuriously increase evidence of linkage. This problem may be of less concern, since direct examination of data for MZ pairs should reveal a surprising degree of genotype identity. Still, given a large study with many families, particularly with substantial missing data, this degree of similarity could be missed, and, in any event, having a means of automatic, rather than manual, detection is desirable. In principle, one can include MZ twins as a possible relationship when using our likelihood ratio method. Given no genotyping error, MZ twins will never be misclassified by our method as full sibs, while, with as few as 22 unlinked markers with four equally frequent alleles, the misclassification rate for full-sib pairs as MZ twins is $<.00000001$. However, in the presence of genotyping error, even one discrepancy between a relative pair formally excludes the possibility of MZ twins. Modifying our method of calculating $P(X|R)$ to explicitly allow for genotyping error would require a much more complicated and computationally demanding approach. We instead recommend one of the IBS-scoring methods for this case, at least when the number of loci tested is not small, with an assessment of whether sharing is significantly greater than that expected when full sibs are assumed.

Although we have described our method to infer relationships in the context of codominant autosomal markers typed on sibships, it is more general. Extensions to allow for X-linked markers or markers demonstrating dominance or to assess other types of noninbred relationships could easily be achieved. Extension to inbred relationships, although not difficult in theory, would require substantially increased computation, since the set of possible IBD relationships between the genes in a relative pair becomes larger. The accuracy of identification of other types of relative pairs, inbred or not, will depend on the true relationship, the other possible relationships, and the degree to which these relationships result in different predictions with regard to the IBD sharing of marker alleles (see Thompson 1975).

In all cases considered, our method resulted in more accurate estimation of pairwise relationships than did the IBS- based methods of Ehm and Wagner (1996) and Stivers et al. (1996). This difference in accuracy was generally modest for evenly spaced markers with equal allele frequencies but, in some cases, became substantial for markers that either had unequal allele frequencies or were not evenly spaced.

It is not surprising that our likelihood ratio method was more accurate than the IBS-scoring methods. Given unequal marker allele frequencies, sharing a rare allele IBS provides stronger evidence for a closer relationship than does sharing a common allele. For example, a relative pair homozygous for an allele with frequency .40 results in a likelihood ratio of 1.75 in favor of full sibs over half sibs, whereas a pair homozygous for an allele with frequency .05 results in a likelihood ratio of 10.50. Our likelihood ratio method makes use of that information, whereas the IBS-scoring methods ignore it. Even given equal marker allele frequencies, simply scoring the alleles IBS is not the same as computing probabilities. For example, given four equally frequent alleles, a 11,11 pair results in a likelihood ratio of 2.50 in favor of full sibs over half sibs, whereas a 12,12 pair results in a likelihood ratio of ∼2.17; the pairs are scored the same by both IBS-scoring methods. Furthermore, our method explicitly allows for linkage of the genetic markers; IBS-scoring methods currently do not. Despite these limitations, the IBS-scoring methods also perform well, given sufficient marker data.

Other approaches to inferring relationships might also be taken. As noted by Goring and Ott (1995), a Bayesian approach is one such alternative. If we believed that full-sib pairs, half-sib pairs, and unrelated pairs occur in sibships in our population with frequencies of a, b, and $c = 1 - a - b$, then we could calculate the posterior probability of a relationship such as full sibs, given the marker data X as

$$P(\text{full sibs}|X) = \frac{aP(X|\text{full sibs})}{\begin{array}{r}aP(X|\text{full sibs}) + bP(X|\text{half sibs})\\ + cP(X|\text{unrelated pairs})\end{array}} .$$

We could then exclude pairs on the basis of small values of $P(\text{full sibs}|X)$.

In conclusion, we have described a method to infer relationships between putative full-sib pairs in a sib-pair linkage study. This method should be valuable to identify non–full-sib pairs in studies in which no other relatives are available. This method could be used either to exclude probable non–full-sib pairs from a linkage study or to separate full-sib pairs and half-sib pairs into two separate samples, with appropriate analysis of each. The method is accurate and easy to perform and requires no laboratory work beyond that which already would be done as part of a genome-scan linkage study. We believe that such an analysis should be a standard component of gene-mapping studies for which sib-pair data without data on additional relatives are to be used.

We have written a FORTRAN 77 program, RELPAIR, that uses the likelihood ratio method to distinguish the most likely relationships between pairs of relatives in a sibship—MZ twins (or duplicated samples), full-sib pairs, half-sib pairs, and unrelated pairs—given data on a set of possibly linked codominant markers. The program is available either on the World Wide Web at http://www.sph.umich.edu/group/statgen/software or by contacting M.B.

Acknowledgments

We thank Margaret Gelder Ehm and Kenneth Lange for helpful discussions and David Stivers for helpful discussions and for making his computer program available to us. We thank Graeme Bell, Pat Concannon, Craig Hanis, and Richard Spielman for permitting reanalysis of their NIDDM genotype data. This research was supported by NIH grants HG00376 (to M.B.) and GM46800, DK20595, DK47481, DK47486, DK47487, and DK47494 (to N.J.C.).

References

Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–98

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

Chakraborty R, Jin L (1993*a*) Determination of relatedness between individuals using DNA fingerprinting. Hum Biol 65:875–895

Chakraborty R, Jin L (1993*b*) A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. In: Pena SDJ, Chakraborty R, Epplen JT, Jeffreys A (eds) DNA fingerprinting: state of the science. Birkhaeuser Verlag, Basel, pp 153–175

Ehm MG, Wagner M (1996) Test statistic to detect errors in sib-pair relationships. Am J Hum Genet Suppl 59:A217

Goring HHH, Ott J (1995) Verification of sib relationship without knowledge of parental genotypes. Am J Hum Genet Suppl 57:A192

Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, et al (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13:161–166

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugenics 12:172–175

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lange K, Boehnke M, Cox DR, Lunetta KL (1995) Statistical methods for polyploid radiation hybrid mapping. Genome Res 5:136–150

Morton NE (1991) Parameters of the human genome. Proc Natl Acad Sci USA 88:7474–7476

Risch N (1990) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253

Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268: 1584–1589

Stivers DN, Zhong Y, Hanis CL, Chakraborty R (1996) RELTYPE: a computer program for determining biological relatedness between individuals based on allele sharing at microsatellite loci. Am J Hum Genet Suppl 59:A190

Thompson EA (1975) The estimation of pairwise relationships. Ann Hum Genet 39:173–188

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–326