Original Paper

European Journal of Human Genetics

Eur J Hum Genet 1997;5:69–77

Received: September 17, 1996
Revision received: December 30, 1996
Accepted: January 21, 1997

Harald H.H. Göring[a]
Jürg Ott[b]

[a] Department of Genetics and Development,
Columbia University, and
[b] Laboratory of Statistical Genetics,
Rockefeller University,
New York, N.Y., USA

Relationship Estimation in Affected Sib Pair Analysis of Late-Onset Diseases

Key Words
Relationship estimation
Identity by descent
Affected sib pair analysis
Late-onset disease

Abstract
In linkage studies, errors in pedigree structure will often be uncovered through Mendelian inconsistencies. In affected sib pair analysis of diseases with late onset, however, such mistakes will usually go undetected since parental genotypes are commonly not known. Cases of nonpaternity, unrecorded adoption or accidental sample swap in the laboratory will then not be noticed. Typically, such relationship errors lead to a decrease in power for linkage. In this paper, a method is presented which allows verification of the relationship between stated sibs using their marker genotypes. The method is likelihood-based and incorporates a Bayesian approach to compute posterior relationship probabilities. It is shown that sibs, half-sibs and unrelated individuals can be distinguished from each other quite reliably using numbers of markers that should be available in most sib pair studies. It is demonstrated that elimination of false sib pairs increases the power to detect linkage in affected sib pair studies. The gain in power may be large if relationship errors occur quite frequently; the gain will be only moderate if relationship errors are very infrequent. Software for relationship estimation is provided.

Introduction

In linkage studies, marker genotypes of pedigree members provide a means to identify mistakes in pedigree structure. Many such errors will be detected during the statistical analysis through occurrence of Mendelian inconsistencies. This is also true for affected sib pair (ASP) analysis if parents are typed or if parental genotypes can be reconstructed from relatives. However, if only two sibs are available and their parents are deceased, which is often the case with late-onset diseases such as Alzheimer's disease or Parkinson's disease, no 'built-in' error control exists. Cases of nonpaternity (or nonmaternity) will then go undetected so that half-sibs will be falsely analyzed as sibs. Similarly, unrecorded adoption, accidental sample swap in the laboratory, inaccurate records or misidentification of individuals will not be noticed. These errors will in most instances lead to the analysis of unrelated individuals as sibs. Since half-sibs and unrelated individuals are expected to share fewer alleles identical by descent (ibd) than sibs such mistakes will typically lead to a loss of power in ASP analysis. Here, relationship estimation based on marker genotypes is proposed to detect non-sibs and thereby increase the power of ASP analysis.

KARGER
E-Mail karger@karger.ch
Fax + 41 61 306 12 34
http://www.karger.ch

Harald Göring
Columbia University, Department of Genetics and Development
722 West 168th Street, Unit 58
New York, NY 10032 (USA)
Tel. (212) 568-9696, Fax (212) 781-2661, E-Mail: hhg2@columbia.edu

Table 1. ibd probabilities for sibs, half-sibs and unrelated individuals

Relationship	k_0	k_1	k_2
Unrelated	1	0	0
Half-sib (also uncle/aunt-nephew/niece, grandparent-grandchild)	1/2	1/2	0
Sib	1/4	1/2	1/4

Absence of inbreeding is assumed. k_a is the probability of having $a = 0$, 1 or 2 alleles ibd at a locus.

Table 2. ibd probabilities at one locus conditional on ibd status at another locus for sibs, half-sibs and unrelated individuals

Relationship	a	$k_{0\|a}$	$k_{1\|a}$	$k_{2\|a}$
Unrelated	0	1	0	0
	1	0	0	0
	2	0	0	0
Half-sib	0	ψ	$1-\psi$	0
	1	$1-\psi$	ψ	0
	2	0	0	0
Sib	0	ψ^2	$2\psi(1-\psi)$	$(1-\psi)^2$
	1	$\psi(1-\psi)$	$\psi^2+(1-\psi)^2$	$\psi(1-\psi)$
	2	$(1-\psi)^2$	$2\psi(1-\psi)$	ψ^2

Absence of inbreeding is assumed. $k_{b|a}$ is the probability of having $b = 0$, 1 or 2 alleles ibd at one locus given a alleles ibd at another locus and inter-locus recombination fraction ϑ. $\psi = \vartheta^2 + (1-\vartheta)^2$.

Theory of Pairwise Relationship Estimation

As a direct consequence of the Mendelian rules of inheritance, genotypes of relatives are similar to each other. The closer the relationship, the higher the degree of genotypic (and phenotypic) similarity. Most methods of relationship estimation require variance computation of genome-wide ibd proportions for different relationships [e.g. ref. 1–4]. Others are based on DNA fingerprinting [5]. The approach taken here is likelihood based and uses marker genotypes. It is based on the theory developed by Thompson [6–8], which is extended to incorporate linked loci. Bayes rule is used to compute posterior relationship probabilities. Even though this method is applicable to all pairwise relationships (i.e. the relationship between two individuals), the focus in this paper will be on sibs, half-sibs and unrelated individuals. Mérette and Ott [9] recently discussed a similar approach using only unlinked loci to estimate the relationship between parents in linkage analysis of recessive traits. Ehm and Wagner [10] recently proposed a test statistic based on identity by state to detect errors in sib-pair relationships.

Unlinked Loci

Assuming Hardy-Weinberg equilibrium, the probability of the joint genotype of two individuals (here called pairwise genotype) at a locus only depends on the relationship between the individuals and on the allele frequencies. This probability may be calculated by conditioning on the number of alleles that are ibd at that locus between the two individuals. Let P_a denote the probability of the pairwise genotype given that the individuals have $a = 0$, 1 or 2 alleles ibd. These conditional probabilities are then independent of the relationship and were tabulated by Thompson [6]. Let k_a denote the probability that two individuals of a certain relationship have $a = 0$, 1 or 2 alleles ibd at a locus (table 1) [6, 11]. The probability of the observed pairwise genotype is then given by

$$P = k_0P_0 + k_1P_1 + k_2P_2$$
$$= \sum_{a=0}^{2} k_aP_a$$

This is a weighted average: the relationship-independent genotype probabilities are weighted by the relationship-specific ibd probabilities. Because of independence of unlinked loci, the joint probability for multiple unlinked loci is given by the product of the probabilities for each locus.

Linked Loci

Linked loci are not independent, so that the product rule for unlinked loci does not apply. Instead, the probability of multiple linked loci must be computed jointly. Let $P^{(i)}_{a_i}$ stand for the probability of the pairwise genotype at locus i (where the superscript denotes the locus, not an exponent) given the individuals have $a_i = 0$, 1 or 2 alleles ibd at that locus. Let the probability of having a_i alleles ibd at locus i conditional on a_{i-1} alleles ibd at locus $i-1$ be denoted by $k^{(i)}_{a_i|a_{i-1}}$ (table 2). These conditional ibd probabilities not only depend on the relationship, but also on the recombination fraction between the two loci. They were given for sibs by Haseman and Elston [12] and derived for other relationships by Bishop and Williamson [13]. Assuming absence of interference, this method can easily be extended to multiple linked loci by conditioning for each locus on the ibd status of the nearest neighboring locus (on the left or on the right, but always on the same side for all linked loci). The probability, P_{LG}, of the pairwise genotypes in a *L*inkage *G*roup consisting of m linked loci is then given by

$$P_{LG} = \sum_{a_1=0}^{2} \sum_{a_2=0}^{2} \dots \sum_{a_m=0}^{2} k_{a_1}^{(1)} P_{a_1}^{(1)} \prod_{i=2}^{m} \left(k_{a_i|a_i-1}^{(i)} P_{a_i}^{(i)} \right)$$

$$= \sum_{a_1=0}^{2} \sum_{a_2=0}^{2} \dots \sum_{a_m=0}^{2} k_{a_1}^{(1)} \prod_{i=2}^{m} k_{a_i|a_i-1}^{(i)} \prod_{i=1}^{m} P_{a_i}^{(i)}.$$

As in the case of an unlinked locus, this probability is again a weighted average: the products of the pairwise genotype probabilities, which are independent of the relationship, are now weighted by the products of the relationship-specific conditional ibd probabilities. This formula also holds for unlinked loci, but is then computationally inefficient. As in the case for unlinked loci (an unlinked locus may be considered a linkage group of its own), the probabilities for different linkage groups are simply multiplied together to give the overall probability of the observed pairwise genotypes for a given relationship. The overall probability for all linkage groups then represents the likelihood for the relationship under which it was computed.

Presence of a Typed Parent

The theory of pairwise relationship estimation described above is based only on the genotypes of the two individuals whose relationship is to be estimated. In ASP analysis, one may still want to verify the relationship between two stated sibs if one parent is typed. Even though Mendelian inconsistencies will now typically lead to detection of unrelated individuals, the two individuals may still be half-sibs rather than sibs. The genotype of the typed parent should then be taken into account in relationship estimation. The probability of the joint genotype of the two individuals is now computed conditional on the observed genotype of the typed parent. The ibd contribution from the typed parent is taken into account separately from that of the untyped parent (true relationship is sib) or parents (true relationship is half-sib). Let k_a and k_b denote the probability of having a and b alleles ibd from the typed parent and the untyped parent(s), respectively. Since sibs can be viewed as having two half-sib contributions (one from each parent), the values in tables 1 and 2 given for half-sibs apply to k_a, and those for either half-sibs or unrelated individuals apply to k_b depending on whether the true relationship is sib or half-sib. Then,

$$P = \sum_{a=0}^{1} \sum_{b=0}^{1} k_a k_b P_{a,b}.$$

Here, $P_{a,b}$ is the probability of the pairwise genotype given the genotype of the typed parent and the ibd contributions of the typed parent and the untyped parent(s). The computation of $P_{a,b}$ involves reconstructing which alleles in the offspring are from the typed parent. If no unambiguous reconstruction is possible, all possible cases need to be considered and weighted appropriately. Transmission probabilities need to be taken into account. In essence, relationship estimation in this situation is reduced to computing whether those alleles in the offspring not coming from the typed parent are more likely to have come from one person (in which case the two individuals are sibs) or from two different persons (in which case the two individuals are half-sibs). The method can be extended for use with linked loci in the same fashion as described above for the case when no parent is typed.

Prior and Posterior Relationship Probabilities

In many situations where an investigator would like to estimate the relationship between two individuals, the prior probabilities of the potential relationships are known to be different. The prior probabilities should then be incorporated in the relationship estimation. Without incorporation of appropriate prior probabilities, the computed (posterior) relationship probabilities are not valid as shown by Elston [14] in the context of paternity testing. Goldgar and Thompson [15] also discussed a Bayesian approach to paternity testing. Let this prior probability for relationship i be denoted by $P(r_i)$ and the corresponding posterior probability (after observing pairwise genotypes g) by $P(r_i|g)$. $P(g|r_i)$ shall stand for the relationship-specific genotype probability computed in the manner described above. Then, by Bayes' rule,

$$P(r_i|g) = \frac{P(g|r_i)P(r_i)}{\sum_j P(g|r_j)P(r_j)},$$

where the denominator is the sum over all possible relationships. The relationship with the highest posterior probability is the estimated relationship.

The difficulty in incorporating prior probabilities lies in assigning appropriate values to them. In ASP analysis, the proportions of half-sibs and unrelated individuals will typically be quite small, but the values are difficult to estimate accurately. These proportions depend on the population under study, on the examined disease as well as on the laboratories involved in the study. Errors in genotyping have been shown to occur quite frequently (>1% [16]) and often appear to be caused by sample swapping [16] which typically results in two unrelated individuals falsely being taken to be sibs. Unrecorded adoption, incorrect records and misidentification of individuals are probably quite rare in most instances. Below, for the case when no parent is typed, we assume a prior probability of 0.01 that two stated sibs are in reality unrelated. If one parent is available for marker typing, this prior probability is assumed to be 0 since such cases would most likely have already been detected by inconsistencies. The frequency of nonpaternity (as well as nonmaternity) presumably varies widely between different cultures. While many societies, including most Caucasian populations, appear to have a fairly low nonpaternity rate (~1% in Switzerland) [17 and references therein], it is probably much higher in other societies [7; see section 3.6: An Amerindian genealogy]. Below, for each of the two cases (zero or one parent typed), we consider a prior probability of 0.01 that two stated sibs are in reality half-sibs. If one parent is typed, the investigator may want to assign different values for half-sib prior probabilities depending on whether the father or mother is available for marker typing. The possibility that two stated sibs are of a relationship other than sib, half-sib or unrelated was not considered here, mainly because this will occur only infrequently. Many other relatives should nonetheless be detectable by the method outlined here, though perhaps with lower probability. While it is difficult to estimate prior probabilities precisely, their influence on posterior probabilities decreases with an increase in the number of typed markers so that no high degree of accuracy in the estimates is required if many markers are typed.

Relationship Estimation in Sib Pair Studies

In this paper, the goal of relationship estimation is not to minimize misclassification between relationships but to maximize power for linkage. Two different approaches are examined below. In one approach (referred to as *weighting approach*) the posterior relationship probabilities are used as weights for affected sib pair and half-sib pair analysis: each stated sib pair in the pool of ascertained affected sib pairs is analyzed for ibd sharing assuming that the true relationship is sib. Each pair is also analyzed assuming that the true relationship is half-sib. The values from both analyses are then weighted by the corresponding posterior relationship probabilities and summed to give the number of alleles shared and nonshared for each pair. These allele-sharing values can then be used to compute the sib pair statistic of choice.

While the weighting approach is straightforward from a theoretical standpoint, it may not be of practical use with existing software for sib pair analysis. We therefore also considered an alternative approach (referred to as *decision approach*), which is more easily employed in practice. In this approach, a decision, based on computed posterior relationship probabilities, is made whether to keep a stated sib pair in the sib pair pool or whether to discard it. The remaining (i.e. nondiscarded) sib pairs are then analyzed as sibs just as in any sib pair analysis.

The following rationale was used to find a suitable decision rule about when to discard a putative sib pair. Since the goal of relationship estimation in this investigation is to maximize the power for linkage, the decision rule should be based on one's expectation, judging solely by the computed posterior relationship probabilities, whether a stated sib pair will or will not provide evidence for linkage if in fact linkage exists. If a pair of individuals is expected to provide linkage evidence, it should be included in the linkage study. In contrast, if a pair of individuals is not expected to provide evidence for linkage, it should be excluded from the linkage study. To make a decision along this reasoning, an investigator may compute the expected proportion of ibd sharing at the (unknown) disease locus as a function of the computed posterior relationship probabilities. If this value is less than 0.5 – which is the expected proportion of ibd sharing for sibs under the null hypothesis of no linkage – a stated sib pair should not be used in the sib pair study, because the pair is not expected to provide linkage evidence, even if linkage exists. Computing this expected value of ibd sharing, however, does require knowledge of disease parameters, which in reality – especially with complex diseases – are often unknown. To derive a general decision rule applicable to most if not all diseases, the maximum rather than the expected proportion of alleles ibd between the 2 individuals is used here. Sibs may share up to 100% their

Table 3. Probabilities of discarding a pair of individuals when no parent is typed

Markers	Markers/ linkage group	Recombination fractions within linkage groups	Probability of discarding		
			unrelateds	half-sibs	sibs
25	1	–	0.944	0.267	0.0008
50	1	–	0.999	0.650	0.0002
50	2	0.100	0.997	0.498	0.0004
50	2	0.010	0.999	0.535	0.0004
100	2	0.100	1.000	0.926	0.0001
100	2	0.010	1.000	0.944	0.0000
100	4	0.100	1.000	0.840	0.0002
100	4	0.010	1.000	0.846	0.0004
200	4	0.100	1.000	0.995	0.0000
200	4	0.010	1.000	0.994	0.0000
200	8	0.100	1.000	0.987	0.0001
200	8	0.010	1.000	0.941	0.0005

Results are based on 1,000 replicates for non-sibs and 10,000 for sibs. All loci have heterozygosity of roughly 0.75 (allele frequencies 0.32, 0.3, 0.2, 0.1, 0.05, 0.02 and 0.01). Prior probabilities for sibs, half-sibs and unrelated individuals are 0.98, 0.01 and 0.01, respectively.

alleles ibd at a locus, half-sibs only up to 50% and unrelated individuals 0% (assuming absence of linkage disequilibrium). These values are then weighted by the computed posterior relationship probabilities. As an example, if the posterior relationship probabilities for sibs, half-sibs and unrelated individuals are 0.7, 0.2 and 0.1, the maximum possible proportion of ibd sharing is 0.7 * 1 + 0.2 * 0.5 + 0.1 * 0 = 0.8. If this maximum possible proportion of ibd sharing is less than 0.5, the putative sib pair should be discarded. This maximum ibd proportion will be obtained only in the (ideal) case that the disease is Mendelian, rare, fully penetrant, recessive and without phenocopies and when the marker is located exactly at the disease and is fully informative. Of course, this situation rarely – if ever – occurs in reality. With most diseases, this maximal ibd sharing is not observed because (1) many disease loci have some dominance component, (2) phenocopies will often be present for heterogeneous diseases, (3) it is thought that several disease loci jointly determine affection with complex diseases, so that allele sharing between affected individuals is often reduced at any particular disease locus, (4) disease alleles may be quite frequent and (5) a marker may not be available right at the disease locus. Discarding stated sibs only for values less than 0.5 will therefore be extremely conservative (i.e. fewer nonsibs will be discarded than desired) in most instances. Here, we chose a maximum ibd proportion of 0.6 as the criterion, which we believe to be still conservative for almost all diseases. If the computed maximum possible proportion of ibd sharing is less than 0.6, a stated sib pair was discarded.

Properties of the Decision Approach

The properties of the decision approach were evaluated via computer simulation on the basis of the following two statistics: (1) The probability that a true sib pair is falsely discarded (smaller is better), and (2) the probability that nonsibs are correctly discarded (larger is better). All simulations in this paper were carried out assuming marker heterozygosities of roughly 0.75 (allele frequencies 0.32, 0.3, 0.2, 0.1, 0.05, 0.02 and 0.01).

When no parent is typed (table 3), unrelated individuals are excluded from subsequent ASP analysis with high probability (>0.9 for as few as 25 unlinked markers). The probability with which half-sibs are discarded is lower (<0.3 for 25 unlinked markers), but increases quickly with additional markers (>0.8 for 100 markers). The probability of falsely discarding sibs is very small (<0.001 for 25 unlinked markers) and decreases further with an increase in the number of markers. When one parent is typed (table 4), sibs and half-sibs can be distinguished more easily than when no parent is typed; the probability of correctly discarding half-sibs is higher (>0.6 for 25 and >0.9 for at least 50 markers, respectively), and the probability of

Table 4. Probabilities of discarding a pair of individuals when one parent is typed

Markers	Markers/ linkage group	Recombination fractions within linkage groups	Probability of discarding	
			half-sibs	sibs
25	1	–	0.614	0.0005
50	1	–	0.948	0.0003
50	2	0.100	0.932	0.0002
50	2	0.010	0.968	0.0000
100	2	0.100	1.000	0.0000
100	2	0.010	1.000	0.0000
100	4	0.100	0.999	0.0000
100	4	0.010	0.999	0.0000
200	4	0.100	1.000	0.0000
200	4	0.010	1.000	0.0000
200	8	0.100	1.000	0.0000
200	8	0.010	1.000	0.0000

Results are based on 1,000 replicates for half-sibs and 10,000 for sibs. All loci have heterozygosity of roughly 0.75 (allele frequencies 0.32, 0.3, 0.2, 0.1, 0.05, 0.02 and 0.01). Prior probabilities for sibs and half-sibs are 0.99 and 0.01, respectively.

falsely discarding true sibs is even lower ($\leq$0.0005 for at least 25 markers). If nonsibs are believed to occur with higher probabilities than assumed here, they will be discarded with even higher probability, while the probability of discarding true sibs falsely is still very low (data not shown).

Power Gain in ASP Analysis through Relationship Estimation

The usefulness of relationship verification between sibs for ASP analysis was assessed by computer simulation. Complete ASP studies of a complex trait were simulated (one ASP study is one replicate), and the p values obtained in the sib pair test with or without prior relationship estimation were recorded for each replicate.

The following disease model was used in the simulations: A quantitative trait with heritability of 50% was assumed to be the underlying cause of the disease. The major disease locus (no dominance component, frequency of disease-predisposing allele 0.2) contributed 30% to the variance of this quantitative trait, while 4 additional loci each contributed 5%. The remaining variation in the quantitative trait belonged in equal parts (i.e. 25% each) to shared and nonshared environmental effects. Individuals in the upper 5th percentile of the distribution were taken to be affected with the disease (i.e. disease prevalence was 5%). In most studies, investigators appear to allow only for non-shared environmental effects. We find that the shared environment acts as a confounder for linkage (similar to background heritability) thus making linkage harder to detect. Since a model with shared environmental effects probably represents reality more closely, such a model has been used in these simulations.

Between 250 and 400 affected pairs with or without a typed parent were simulated for each replicate. The true relationship of each pair was simulated according to chosen prior relationship probabilities. A marker linked to the major disease locus with recombination fraction 0.01 was simulated (assuming absence of linkage disequilibrium). The ASP test was based on this marker. 100 additional markers located throughout the genome but unlinked to the major disease locus were simulated for each individual. Relationship estimation was based on these 100 markers.

Each replicate was then analyzed in 4 different ways. In method 1, all pairs, irrespective of the true relationship, were included in ASP analysis. This method represents ignorance regarding relationship misspecifications and is the normal way of analysis in most studies. In methods 2 and 3, the relationship between the individuals was estimated (based on the 100 simulated markers). In method 2, the decision approach was followed; in method 3, the weighting approach was used. In method 4, the affected pairs were analyzed under their true relationship, i.e. sibs were analyzed as sibs, half-sibs as half-sibs and unrelated individuals were not analyzed since they provide no linkage information. This last method represents the ideal and hypothetical case without relationship errors. The p values of the mean test of allele sharing [18] were recorded for all four methods of analysis. Each result is based on 1,000 replicates.

Table 5 gives the power of the 4 methods of analysis, that is, the probability of finding a significant linkage result in the ASP test, where significance is defined as achieving an empirical significance level, $p \leq 0.000022$ (corresponding to an asymptomatic lod score of 3.6). This value has been suggested as a suitable threshold for declaring a linkage finding in an affected sib pair study as significant [19]. In addition, the average (as judged by the median over replicates) ratio of the observed p value without relationship estimation and the p values obtained with the different methods of relationship estimation is given in table 5. These ratios indicate by what factor one can expect on average (median) the p value of a sib pair study to decrease when relationship estimation is used.

Table 5. Effect of relationship estimation on ASP studies for a complex trait

Prior relationship probabilities (sib, half-sib, unrelated)	Parents typed	Families		Method of analysis			
				1	2	3	4
0.88, 0.10, 0.02	0	400	power	0.035	0.372	0.381	0.387
			p ratio	1	72.039	77.282	81.459
0.98, 0.01, 0.01	0	350	power	0.277	0.379	0.388	0.392
			p ratio	1	2.560	2.628	2.648
0.90, 0.10, 0	1	300	power	0.109	0.378	0.417	0.417
			p ratio	1	19.921	28.473	28.426
0.99, 0.01, 0	1	250	power	0.267	0.297	0.298	0.298
			p ratio	1	1.277	1.324	1.326

Shown is the power of the simulated ASP study to obtain a significant linkage finding with and without different methods of relationship estimation. Also shown is the median ratio of the p value obtained in the ASP study without relationship estimation and the p values obtained when the different methods of relationship estimation are used. Refer to the text on details about the disease model used and a description of the simulations. Results are based on 1000 replicates.

Method 1: Analysis of all affected pairs as sibs regardless of true relationship.
Method 2: Analysis involving decision approach of relationship estimation.
Method 3: Analysis involving weighting approach of relationship estimation.
Method 4: Analysis under true relationship (no relationship errors).

The results show that relationship estimation can greatly increase the power for linkage in ASP studies if relationship errors are quite frequent. However, if these errors are infrequent, the effect on the ASP study is much smaller. When relationship errors occur very infrequently, relationship estimation is neither necessary nor very useful. As expected, the higher the proportion of non-sibs, the greater the gain in power. With relationship estimation based on 100 markers as in these simulations, both methods involving relationship estimation prior to ASP analysis (i.e. the weighting approach and the decision approach) closely approach the ideal hypothetical situation without relationship errors.

Discussion

The presence of half-sibs and unrelated individuals in a pool of ascertained affected sib pairs will typically lead to a loss of power in ASP analysis. The higher the proportion of non-sibs and the lower the power to detect linkage a priori, as with complex diseases, the more pronounced this effect. The impact of unrelated individuals is severer than that of half-sibs. While nonsibs are typically identified by Mendelian inconsistencies when parental genotypes are known, this is generally not the case when parental genotypes are unavailable. The outlined method of verifying the relationship between stated sibs allows the detection of such relationship errors in most instances if a sufficient number of markers has been typed. This method should therefore be useful for ASP studies of diseases with late age of onset, where parents of affected individuals are often dead. It has been proposed [20] that it may be more efficient, from a standpoint of cost effectiveness in linkage analysis, not to type parents but rather additional affected sib pairs, especially if marker heterozygosity is high (as with most modern markers) and if there is no shortage of affected sib pairs (as with common diseases). Relationship estimation may give the linkage analyst more 'faith' in such an approach.

Throughout this paper, it has been implicitly assumed (in the presented theory as well as the simulations) that male and female recombination fractions are equal. When one parent is typed, differences in recombination rates between the sexes could easily be incorporated into the theory, however. When no parent is typed, this would neither be as straightforward nor as useful.

Absence of inbreeding was assumed in this investigation. If this is not the case, the ibd probabilities given in tables 1 and 2 do not hold, since the presence of inbreeding would lead to increased ibd sharing. If the ibd probabilities from tables 1 and 2 are nonetheless used, inbred sibs as well as inbred half-sibs and unrelated individuals will be discarded with lower probabilities than those shown in tables 3 and 4.

The posterior relationship probabilities are influenced by misspecification of marker allele frequencies. Incorrectly assuming equal allele frequencies at all loci will typically lower the probability to detect non-sibs, while the risk of falsely mistaking true sibs for non-sibs will also be decreased (data not shown). It has been found that incorrectly assuming equal allele frequencies often leads to false-positive evidence for linkage in lod score analysis of pedigrees with untyped individuals [21]. Similarly, allele sharing is biased upwards, which in ASP analysis may lead to false-positive evidence for linkage. In relationship estimation, it results in a bias towards closer relationships. Since relationship estimation is based on many marker genotypes jointly, it is more robust to wrong allele frequencies than sib pair statistics, which are based on only a single marker or a few markers in case of multipoint analysis, when parents are not typed. Allele frequencies should therefore have already been estimated for the sake of sib pair analysis.

Relationship estimation and ASP analysis are both based on ibd sharing distributions. If the same markers are used for both methods, the results are not independent of each other. This may potentially lead to false-positive evidence for linkage in ASP analysis when the relationship between sibs is verified via relationship estimation. As an example, consider the decision approach and assume all stated sibs are true sibs. If any sib pairs are now (falsely) excluded from the linkage study, the remaining sib pairs are no longer a random sample of affected sib pairs; they are depleted of sibs sharing by chance only a small number of alleles ibd (because it is these very sib pairs that were mistaken for nonsibs and discarded). The chance of a false-positive linkage finding is then increased. This problem is completely avoided if two different (and unlinked) sets of markers are used for ASP analysis and relationship estimation. To avoid typing extra markers only for relationship estimation, this may be most easily achieved by only using those markers for relationship estimation which are located in regions other than where the investigator is currently looking for linkage to the disease. Relationship estimation is therefore perhaps most easily employed after encouraging linkage evidence has been found. If many markers throughout the genome are typed, the evidence for a relationship error is often so overwhelming that omitting markers in any genomic region from relationship estimation does not make any difference regarding the decision on discarding.

A bias for relationship estimation may occur if markers that are linked to susceptibility loci are used in relationship estimation. This is the case because the values of ibd probabilities given in tables 1 and 2 would not be true for sibs and half-sibs in that situation – similar to the case where inbreeding is present as discussed above. Using markers linked to disease loci would typically make affected individuals appear more closely related than they are in reality since the genotypes of affected individuals are expected to be similar at markers linked to disease loci. True sibs would be even less likely to be mistaken for non-sibs, but half-sibs would be slightly more difficult to detect. Assuming absence of linkage disequilibrium, this situation would not affect the ability to detect unrelated individuals. Because relationship estimation is based on many markers jointly, this bias on relationship estimation should be small. Its direction is 'safe' with respect to the sib pair study if the guideline from above regarding which markers to use for relationship estimation is followed. For complex traits, the situation is more complicated, and it is important that individuals are only discarded if there is good evidence that two individuals are not sibs.

When fewer than, say, 25 markers are typed or when the available markers fall into only a small number of linkage groups or when linkage groups with tightly linked markers contain very different numbers of markers, relationship estimation is less accurate. In the decision approach, the probability with which true sibs are falsely discarded may on occasion be much higher than shown (data not shown). This may be unacceptable if the suspected proportion of half-sibs and unrelated individuals is low because more true sibs than nonsibs may then be discarded. We do not recommend using relationship estimation in this case.

It may be of interest to compare the information for relationship estimation provided by unlinked and linked markers, but such a comparison depends on the true relationship and the alternative relationships in question. As an extreme example, the relationship half-sib, grandparent-grandchild and uncle/aunt-nephew/niece have identical ibd probabilities for unlinked loci. These three relationships can therefore only be distinguished using *linked* loci [7]. For the three relationships considered here, unlinked loci provide more information (see tables 3 and 4; additional data not shown). In addition, occasional genotyping errors presumably have only a negligible influence on relationship estimation if they occur in unlinked markers. With linked markers, especially if linkage is tight, errors in genotyping will have a greater, yet probably still small, effect, at least as long as the available markers fall into many linkage groups. Even though unlinked markers are preferable, it is clear from the results shown in tables 3 and 4 that, due to the limited size of the genome, linked markers provide valuable information for relationship estimation. It may be tempting not to bother with the recombination fractions between markers and to simply treat all markers as unlinked despite knowledge to the contrary. This, however, is not advisable. While this perhaps may not lead to a systematic bias, relationship estimation becomes much less reliable and true sibs may be falsely mistaken for nonsibs with quite a high frequency (data not shown). Small and unavoidable errors in the estimates of the intermarker recombination fractions should not influence relationship estimation significantly.

Software

Software for relationship estimation (program 'relative'), which accepts normal LINKAGE format input files, is available by anonymous ftp from linkage.rockefeller.edu in subdirectory software/relative.

Acknowledgments

This material is based upon work supported by US Army Medical Research Acquisition Activity under award No. DAMD17-94-J-4406. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the US Army Medical Research Acquisition Activity.

References

1 Suarez BK, Reich T, Fishman PM: Variability in sib pair genetic identity. Hum Hered 1979; 29:37–41.
2 Risch N, Lange K: Application of a recombination model in calculating the variance of sib pair genetic identity. Ann Hum Genet 1979;43: 177–186.
3 Rasmuson M: Variation in genetic identity within kinships. Heredity 1993;70:266–268.
4 Guo S-W: Variation in genetic identity among relatives. Hum Hered 1996;46:61–70.
5 Chakraborty R, Jin L: Determination of relatedness between individuals using DNA fingerprinting. Hum Biol 1993;65:875–895.
6 Thompson EA: The estimation of pairwise relationships. Ann Hum Genet 1975;39:173–188.
7 Thompson EA: Pedigree analysis in human genetics. Baltimore, Hopkins, 1986.
8 Thompson EA: Estimation of relationship from genetic data; in Rao CR, Chakraborty R (eds): Handbook of statistics. North Holland, Amsterdam, 1991, vol 8, pp 255–269.
9 Mérette C, Ott J: Estimating parental relationship in linkage analysis of recessive traits. Am J Med Genet 1996;63:386–391.
10 Ehm MG, Wagner M: Test statistic to detect errors in sib-pair relationships. Am J Hum Genet 1996;59:A217 (abstract).
11 Cotterman CW: A calculus for statistico-genetics; PhD thesis, Ohio State University, 1940; in Ballonoff P (ed): Genetics and Social Structure. Benchmark Papers in Genetics. Dowden, Hutchinson & Ross, 1975, pp 157–272.
12 Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1972;2:3–19.
13 Bishop DT, Williamson JA: The power of identity-by-state methods for linkage analysis. Am J Hum Genet 1990;46:254–265.
14 Elston RC: Probability and paternity testing. Am J Hum Genet 1986;39:112–122.
15 Goldgar DE, Thompson EA: Bayesian interval estimation of genetic relationships: Application to paternity testing. Am J Hum Genet 1988;42:135–142.
16 Brzustowicz LM, Mérette C, Xie X, Townsend L, Gilliam TC, Ott J: Molecular and statistical approaches to the detection and correction of errors in genotype databases. Am J Hum Genet 1993;53:1137–1145.
17 Sasse G, Müller H, Chakraborty R, Ott J: Estimating the frequency of nonpaternity in Switzerland. Hum Hered 1994;44:337–343.
18 Blackwelder WC, Elston RC: A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 1985;2:85–97.
19 Lander E, Kruglyak L: Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nature Genet 1995; 11:241–247.
20 Holmans P: Asymptomatic properties of affected-sib-pair linkage analysis. Am J Hum Genet 1993;52:362–374.
21 Ott J: Strategies for characterization highly polymorphic markers in human gene mapping. Am J Hum Genet 1992;51:283–290.