Genetic Epidemiology 20:307–315 (2001)

Estimation of Allele Frequencies With Data on Sibships

Karl W. Broman*

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland

Allele frequencies are generally estimated with data on a set of unrelated individuals. In genetic studies of late-onset diseases, the founding individuals in pedigrees are often not available, and so one is confronted with the problem of estimating allele frequencies with data on related individuals. We focus on sibpairs and sibships, and compare the efficiency of four methods for estimating allele frequencies in this situation: (1) use the data for one individual from each sibship; (2) use the data for all individuals, ignoring their relationships; (3) use the data for all individuals, taking proper account of their relationships, considering a single marker at a time; and (4) use the data for all individuals, taking proper account of their relationships, considering a set of linked markers simultaneously. We derived the variance of estimator 2, and showed that the estimator is unbiased and provides substantial improvement over method 1. We used computer simulation to study the performance of methods 3 and 4, and showed that method 3 provides some improvement over method 2, while method 4 improves little on method 3. Genet. Epidemiol. 20:307–315, 2001.

Key words: sibpairs; late-onset disease; allele frequencies

INTRODUCTION

Statistical methods to identify genes by linkage or linkage disequilibrium analysis generally require estimates of allele frequencies at genetic markers. Misspecification of marker allele frequencies can lead to increased false-positive rates in linkage analysis [Ott, 1992]. For example, the assumption of equally frequent alleles is usually wrong and can lead to erroneous inferences.

*Correspondence to: Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205. E-mail: kbroman@jhsph.edu

Received 31 July 2000; Accepted 27 September 2000

Allele frequency estimates are generally derived from data on a set of unrelated individuals. In studies of late-onset diseases, however, one is often confronted with data on affected sibpairs or sibships for which parental genotypes are unavailable. Several different approaches have been taken to estimate allele frequencies in such situations. Some studies have used a single individual from each family [e.g., Witte et al., 2000]. Others have used all individuals, ignoring their relationships [e.g., Weeks et al., 2000].

Boehnke [1991] described the computation of maximum likelihood estimates (MLEs) of allele frequencies, taking proper account of the relationships between individuals (implemented in the USERM13 module of MENDEL [Lange et al., 1988]). In an application of this method to data on a set of 233 individuals in 52 small families, typed at a single polymorphic marker, Boehnke showed that an important improvement in the allele frequency estimates may be obtained by the appropriate use of data on all individuals. The appropriateness of ignoring the relationships between individuals, and the gain obtained by taking proper account of their relationships, remained an open question.

We reconsider the problem of estimating allele frequencies with data on related individuals; we focus on the case of sibpairs or sibships for which parental genotypes are not available. We compare the efficiency of four methods for estimating allele frequencies: (1) use the data for one individual from each sibship; (2) use the data for all individuals, ignoring their relationships; (3) use the data for all individuals, taking proper account of their relationships, considering a single marker at a time; and (4) use the data for all individuals, taking proper account of their relationships, considering a set of linked markers simultaneously. While methods 1 and 2 may be studied analytically, estimators 3 and 4 cannot be written in closed form (we use a version of the EM algorithm [Dempster et al., 1977]), and so their properties must be studied via computer simulation.

We derived the variance of estimator 2 and showed that it is unbiased and provides a substantial improvement over method 1; in the case of sibships of varying sizes, it is best to combine family-specific estimates, using weights inversely proportional to the variances. The results of computer simulations showed that method 3 provides some improvement over method 2, though at the cost of an increase in computational effort, while method 4 improves little on method 3 and requires a great increase in computation.

METHODS AND RESULTS

We seek to estimate the population allele frequencies for an autosomal marker, with genotype data on siblings, in the case that parental genotypes are not available. Consider n sibships, and let k_i denote the number of siblings in family i.

We assume Hardy-Weinberg equilibrium and random mating, that the $2n$ parents are unrelated, and that genotyping errors are absent. In the consideration of multiple markers (method 4 below), we assume that the markers are in linkage equilibrium, that the recombination process exhibits no interference and no sex difference, and that the genetic map is known exactly. Without loss of generality, we will focus on estimating the frequency of allele 1 at a single marker. Let p denote its true underlying frequency, and let X_{ij} be the number of 1 alleles carried by sibling j in family i (= 0, 1, or 2).

Note that $EX_{ij} = 2p$, $\text{var}X_{ij} = 2p(1-p)$, and $\text{cov}(X_{ij}, X_{ij'}) = 4\phi_{ijj'}p(1-p)$, where $\phi_{ijj'}$ is the kinship coefficient for individuals j and j' in family i. (For a given relative pair, $\phi = \pi_1/4 + \pi_2/2$, where π_k is the probability that the pair share k alleles identical by descent (IBD) at an autosomal locus.) The calculation of $\text{cov}(X_{ij}, X_{ij'})$ follows relatively simply, after conditioning on the IBD status of the pair (see the Appendix).

Method 1: One Sibling Per Family

Consider the estimate $\hat{p}^{(1)} = \Sigma_{i=1}^{n} X_{i1}/(2n)$, based on data from the first (or a randomly chosen) sibling from each family. This is an estimate based on n unrelated individuals, and so is unbiased and has variance $p(1-p)/(2n)$. Below, we evaluate the relative efficiency of estimators that make use of additional data, compared to this first method.

Method 2: All Individuals, Ignoring Relationships

Consider the estimated allele frequency based on data for the ith family, ignoring their relationships: $\hat{p}_i^{(2)} = \Sigma_{j=1}^{k_i} X_{ij}/(2k_i)$. It is easy to show that this estimate is unbiased and has variance $\bar{\phi}_i p(1-p)$ (see the Appendix), where $\bar{\phi}_i$ is the average kinship coefficient for family i, where the average is over all k_i^2 pairs of individuals (including each individual with himself): $\bar{\phi}_i = \Sigma_{j,j'}\phi_{ijj'}/k_i^2$.

For a sibship, since $\phi_{ijj'} = 1/2$ for an individual and $\phi_{ijj'} = 1/4$ for a sibpair, $\bar{\phi}_i = (k_i + 1)/(4k_i)$. Thus, the variance of $\hat{p}_i^{(2)}$ is $p(1-p)(k_i + 1)/(4k_i)$, and the relative efficiency of this estimate (compared to method 1, using a single individual) is $2k_i/(k_i + 1)$. For a sibpair ($k_i = 2$), the relative efficiency is 4/3. As $k_i \to \infty$, the relative efficiency approaches 2 (since with a very large sibship, one will be able to infer the 4 parental alleles).

In combining data across n families, one may simply group all individuals together, to obtain

$$\hat{p}^{(2)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i} X_{ij}}{2\sum_{i=1}^{n} k_i} = \frac{\sum_{i=1}^{n} \hat{p}_i^{(2)} k_i}{\sum_{i=1}^{n} k_i}.$$

This is equivalent to combining the $\hat{p}_i^{(2)}$ with weights k_i.

An improved estimator is obtained by combining the family-specific estimates $\hat{p}_i^{(2)}$ with weights inversely proportional to the variances, i.e., $k_i/(k_i + 1)$. This is equivalent to combining the X_{ij} with weights $1/(k_i + 1)$, and gives

$$\hat{p}^{(2')} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i} X_{ij}/(k_i+1)}{2\sum_{i=1}^{n} k_i/(k_i+1)}.$$

The relative efficiency of $\hat{p}^{(2)}$, vs. $\hat{p}^{(1)}$, is $2(\Sigma_i k_i)^2/[n\Sigma_i k_i(k_i + 1)]$, while that for $\hat{p}^{(2')}$, is $2[\Sigma_i k_i/(k_i + 1)]/n$. The latter is always at least as big as the former; the two are

the same when the k_i are constant (i.e., all sibships are the same size), in which case the relative efficiency is $2k/(k + 1)$. Note that both of these estimators are unbiased for p.

For example, consider 25 sibpairs and 25 sibtrios. In forming $\hat{p}^{(2)}$, the X_{ij} for each of the 125 individuals are given equal weight, and the resulting estimator has a relative efficiency (vs. $\hat{p}^{(1)}$) of ~1.39. In forming $\hat{p}^{(2')}$, the X_{ij} for the individuals from sibpairs are given weight 1/3, while the X_{ij} for the individuals from sibtrios are given somewhat lesser weight, 1/4. The resulting estimator has a relative efficiency (vs. $\hat{p}^{(1)}$) of ~1.42.

Method 3: Accounting for Relationships

Boehnke [1991] described a general approach to computing MLEs of allele frequencies, accounting for the relationships between individuals. We describe two alternative approaches, specific for the cases of sibpairs and sibships, which are less general but provide some improvement in computation time. We make use of the EM algorithm [Dempster et al., 1977].

For data on sibpairs, we count the observed allele frequencies in the two siblings, but weight each allele in the second sibling by the estimated probability that it is not IBD with one of the first sibling's alleles. An allele in the second sibling that is distinct from either of the first sibling's alleles is given weight 1, since it cannot be IBD with one of the first sibling's alleles. If the allele (say it is allele l) is identical by (or in) state (IBS) with one of the first sibling's alleles, the probability that it is not IBD with one of the first sibling's alleles is $p_l/(1 + p_l)$ (where p_l is the frequency of allele l), except in the case that the second sibling is homozygous while the first sibling is heterozygous, in which case the probability is $(1 + p_l)/(2 + p_l)$.

One begins with initial estimates of the allele frequencies (obtained, for example, by method 2, above). New estimates are derived, giving full weight to the alleles in the first sibling from each pair and weighting the alleles in the second sibling from each pair by the probabilities described above, with the allele frequencies p_l replaced by their current estimates. This process is iterated until the estimates converge.

In the case of data on larger sibships, we use a somewhat more complex form of the EM algorithm. We include the parental genotypes in the augmented data. As a result, the sufficient statistics are the counts of parental alleles, and we base the allele frequencies on the expected numbers of alleles in the parents given the observed data on the children. For each sibship, we sum over all possible parental genotypes g_1, g_2, and calculate the allele frequencies in the parents, weighting the alleles by $\Pr(g_1,g_2|\boldsymbol{c})$, which is proportional to $\Pr(g_1)\Pr(g_2)\ \prod_j \Pr(c_j|g_1,g_2)$, where c_j denotes the genotype of child j. We again start with some initial estimates, replace (in the weights) the allele frequencies with their current estimates, and iterate until convergence. For markers with many alleles, this procedure is quite time consuming, due to the large number of possible parental genotypes.

Because these estimators cannot be written in closed form, analytic derivation of their variances is not feasible. We studied their performance by computer simulation (described below).

Method 4: Multipoint Estimate

The use of data on a set of linked markers provides improved estimates of the number of alleles shared IBD between two individuals, and so such information may

be used to obtain improved estimates of allele frequencies. We applied this multipoint approach to the special case of sibpairs.

We followed the approach, for sibpairs, of method 3 (described above), in which each allele in the second sibling is weighted according to the estimated probability that it is not IBD with either of the first sibling's alleles. Here, the probabilities (weights) were calculated conditional on the genotype data for all markers (assuming no crossover interference). We used the hidden Markov model (HMM) technology developed by Baum et al. [1970], as described in Boehnke and Cox [1997]. The computational effort required to obtain allele frequency estimates increases tremendously, because within each iteration to update the allele frequency estimates, the forward/backward equations for the HMM must be used to re-estimate the weights.

Computer Simulations

While we could derive, analytically, the variances of estimators 1 (one sibling per family) and 2 (ignoring relationships), the performance of methods 3 (accounting for relationships) and 4 (multipoint estimate) needed to be studied by computer simulations. We considered two cases.

First, we simulated the genotype data for 100 sibpairs at a set of 10 linked markers, with equal inter-marker distances of d cM, where d was taken to be 1, 5, 10, or 15. Each marker had at least five alleles. The frequencies of the first four alleles were 0.05, 0.10, 0.15, and 0.20. The number and frequencies of other alleles were chosen to give a marker heterozygosity (het) of 0.70, 0.75, 0.80, 0.85, or 0.90. For each of the 20 cases (4 values for $d \times 5$ values for het), 10,000 replicates were performed, and each of methods 1–4 were applied to estimate the frequencies of the first four alleles. All methods were found to be approximately unbiased. The estimated variances of estimators 1 and 2 corresponded closely to the values calculated analytically. Recall that $\text{var}(\hat{p}^{(1)}) = p(1-p)/(2n)$ and that the relative efficiency (in the case of sibpairs) of $\hat{p}^{(2)}$ vs. $\hat{p}^{(1)}$ (i.e., $\text{var}\hat{p}^{(1)}/\text{var}\hat{p}^{(2)}$) is 4/3.

The estimated relative efficiency of $\hat{p}^{(3)}$ (accounting for relationships) vs. $\hat{p}^{(1)}$ is shown in Table I. (Note that, because this approach is based on data for one marker at a time, these estimates are obtained after combining results across marker locations and values of the intermarker distance d. Thus, each entry in the table is derived from $10 \times 4 \times 10{,}000 = 400{,}000$ replicates.) The standard errors for the values in Table I are all < 0.01. While for $\hat{p}^{(2)}$ the relative efficiency is constant in the allele

TABLE I. Estimated Relative Efficiency of $\hat{p}^{(3)}$ (Accounting for Relationships, One Marker at a Time) vs. $\hat{p}^{(1)}$ (One Sibling Per Family) for Data on 100 Sibpairs, as a Function of the Allele Frequency (p) and Marker Heterozygosity (het)

	p			
het	0.05	0.10	0.15	0.20
0.70	1.45	1.44	1.43	1.42
0.75	1.45	1.44	1.44	1.42
0.80	1.45	1.45	1.44	1.43
0.85	1.47	1.46	1.46	1.44
0.90	1.48	1.47	1.46	1.46

frequency p, for $\hat{p}^{(3)}$ the relative efficiency is slightly larger for smaller values of p. The relative efficiency also shows a modest increase as the marker heterozygosity increases. In all cases $\hat{p}^{(3)}$ (accounting for relationships) shows some improvement over $\hat{p}^{(2)}$.

The estimated relative efficiency of $\hat{p}^{(4)}$ (multipoint estimate) vs. $\hat{p}^{(1)}$ is shown in Table II. (Similar results were obtained for each of the 10 marker positions, and so the entries in Table II are obtained after combining results across markers. Each entry is derived from 100,000 replicates.) Here, marker heterozygosity shows only a small effect, and only when the inter-marker distance (d) is small and the allele frequency (p) is large. Inter-marker distance also has a slight effect, but only when the allele frequency is large. In comparing the results to those of Table I, one sees that $\hat{p}^{(4)}$ provides little improvement over $\hat{p}^{(3)}$.

In order to study the efficiency of $\hat{p}^{(3)}$ in the case of larger sibships, we performed a second set of computer simulations, with data on 100 families of 2–6 siblings, with average size 3.5. (The numbers of families with 2, 3, 4, 5, and 6 siblings were 24, 31, 26, 14, and 5, respectively.) We performed 100,000 replicates of a single marker, with marker allele frequencies as in the previous simulations. We did not apply the multipoint approach for these simulations.

The relative efficiency $\hat{p}^{(2)}$, $\hat{p}^{(2')}$, and $\hat{p}^{(3)}$ vs. $\hat{p}^{(1)}$ are displayed in Table III. Marker heterozygosity (het) was found to have little effect on the relative efficiency of $\hat{p}^{(3)}$ in this situation, and so the results in the last row of Table III are averaged across the values of het considered. Note that the values for $\hat{p}^{(2)}$ and $\hat{p}^{(2')}$ are based on analytical calculations and do not depend on the allele frequency, p. An appropriate weighting of different sizes ($\hat{p}^{(2')}$) provides some improvement over ignoring the family structures completely ($\hat{p}^{(2)}$). As before, considerable improvement is gained when accounting for the relationships between individuals, especially for rare alleles.

TABLE II. Estimated Relative Efficiency of $\hat{p}^{(4)}$ (Multipoint Estimate) vs. $\hat{p}^{(1)}$ (One Sibling Per Family) for Data on 100 Sibpairs, as a Function of the Allele Frequency (p), Inter-Marker Distance (d, in cM), and Marker Heterozygosity (het)

		p			
het	*d*	0.05	0.10	0.15	0.20
0.7	1	1.48	1.47	1.46	1.45
	5	1.50	1.45	1.44	1.42
	10	1.47	1.45	1.42	1.41
	15	1.47	1.44	1.42	1.40
0.8	1	1.49	1.48	1.46	1.44
	5	1.48	1.45	1.44	1.43
	10	1.47	1.44	1.43	1.41
	15	1.48	1.44	1.43	1.41
0.9	1	1.50	1.49	1.47	1.48
	5	1.48	1.46	1.46	1.45
	10	1.48	1.44	1.44	1.41
	15	1.48	1.46	1.43	1.42

TABLE III. Estimated Relative Efficiency of $\hat{p}^{(2)}$ (Ignoring Relationships), $\hat{p}^{(2')}$ (Ignoring Relationships; Weighting Families Appropriately) and $\hat{p}^{(3)}$ (Accounting for Relationships, One Marker at a Time) vs. $\hat{p}^{(1)}$, for Data on 100 Sibships of Varying Size (Average No. Siblings = 3.5), as a Function of the Allele Frequency (p)

	p			
Method	0.05	0.10	0.15	0.20
2		1.43		
2′		1.52		
3	1.71	1.69	1.68	1.67

DISCUSSION

In estimating allele frequencies with data on related individuals, one will not go far wrong in simply using the data for all individuals, ignoring their relationships. The resulting estimator is unbiased, improves on the use of one individual from each family, and, importantly, has a variance that is easily calculated. Of course, if data on the founding individuals in pedigrees are available, one should make use of them. (For example, one may compare the variance of the estimator based on only the founders to that based on all family members, ignoring their relationships. In the case of a sibship with one available parent, the latter estimator has smaller variance, whereas if both parents are available, the former has smaller variance.) In addition, it is best to at least partly consider family structure, obtaining separate allele frequency estimates for each family and then combining these, using weights inversely proportional to the variances. Estimates based on a single individual from each family have an additional disadvantage: alleles present in the data may have frequency estimates of 0, if the relevant individuals were not among those included in forming the estimates.

While properly accounting for the relationships between individuals can lead to considerably improved estimates of allele frequencies, this improvement may be of little consequence in situations in which the allele frequencies are not of central interest but rather are nuisance parameters (as is generally the case in linkage analysis). For example, consider an allele with frequency 0.10. The standard error (SE) of the frequency estimate, based on 100 unrelated individuals, is ~0.021. The estimate based on 100 sibpairs, ignoring their relationships, has SE $\approx$ 0.018 (equivalent to adding data on an additional 33 unrelated individuals). The estimate based on 100 sibpairs, properly accounting for their relationships, also has SE $\approx$ 0.018. In this situation, accounting for the relationships between sibpairs is essentially equivalent to adding data on an additional 11–13 unrelated individuals; such a slight improvement may not be worth the additional computational effort.

The multipoint approach to estimating allele frequencies is likely to be solely of academic interest. Some readers may be entertained by the notion of using linkage to estimate allele frequencies, but the great increase in computation is not worth the only marginal improvement in the estimates. Because a sibpair shows, on average, three distinct alleles at an autosomal marker, the ideal estimated allele frequency would have relative efficiency (with respect to the estimate based one sibling from each pair) of 1.5. The values in Table I are quite close to this upper limit.

It is important to point out that the above results rely on the assumption that the subjects were randomly ascertained, which is seldom true in a genetic study. Con-

sider the more common case of affected sibpairs. For a marker that is in linkage disequilibrium with a disease susceptibility gene, the estimated allele frequencies may be biased, even if one uses a set of unrelated affected individuals. If the marker is in equilibrium with any disease genes, the estimated allele frequencies will be unbiased, whether based on one or several siblings from each family.

If the marker is linked to a disease susceptibility locus, then the IBD probabilities for an affected sibpair, at the marker, will deviate from the null probabilities, and as a result, the estimated allele frequencies will have larger variances. In the extreme case of complete sharing at the marker, the inclusion of data on the second sibling will provide no improvement in the allele frequency estimates, but the estimates cannot deteriorate in quality.

ACKNOWLEDGMENTS

Melanie Bahlo, Terri Beaty, Michael Boehnke, and Yin Yao Shugart generously provided comments for improvement of the manuscript.

REFERENCES

Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41:164–71.

Boehnke M. 1991. Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–5.

Boehnke M, Cox NJ. 1997. Accurate inference of relationships in sib-pair linkage studies. Am J Hum Genet 61:423–29.

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

Lange K, Weeks D, Boehnke M. 1988. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471–2.

Ott J. 1992. Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet 51:283–90.

Weeks DE, Conley YP, Mah TS, Paul TO, Morse L, Ngo-Chang J, Dailey JP, Ferrell RE, Gorin MB. 2000. A full genome scan for age-related maculopathy. Hum Mol Genet 9:1329–49.

Witte JS, Goddard KAB, Conti DV, Elston RC, Lin J, Suarez BK, Broman KW, Burmester JK, Weber JL, Catalona WJ. 2000. Genomewide scan for prostate cancer-aggressiveness loci. Am J Hum Genet 67:92–9.

APPENDIX

Consider a noninbred relative pair. Let π_k denote the probability that the pair share k alleles identical by descent (IBD). Let $\phi = \pi_1/4 + \pi_2/2$ denote the kinship coefficient for the pair. We focus on allele 1 at a single autosomal marker. Let p denote its population frequency, and let X_1, X_2 denote the number of 1 alleles carried by the two individuals. We week to calculate $\text{cov}(X_1, X_2)$, under the assumption of Hardy-Weinberg equilibrium and random mating. Note that $\text{cov}(X_1, X_2) = \text{E}(X_1, X_2) - \text{E}(X_1)\text{E}(X_2) = \text{E}(X_1 X_2) - 4p^2$.

We need to find the joint distribution of X_1 and X_2. We condition on the IBD status of the relative pair, to obtain $\Pr(X_1 = i, X_2 = j | \text{IBD} = k)$, displayed in Table IV. From Table IV, we may calculate, with a simple though somewhat tedious bit of algebra,

TABLE IV. Joint Distribution of the Numbers of 1 Alleles Carried by Two Individuals, Given the Number of Alleles They Share IBD

IBD	X_1, X_2	$\Pr(X_1, X_2 \mid \text{IBD})$
0	0,0	$(1-p)^4$
	0,1	$2p(1-p)^3$
	1,0	$2p(1-p)^3$
	1,1	$4p^2(1-p)^2$
	0,2	$p^2(1-p)^2$
	2,0	$p^2(1-p)^2$
	1,2	$2p^3(1-p)$
	2,1	$2p^3(1-p)$
	2,2	p^4
1	0,0	$(1-p)^3$
	0,1	$p(1-p)^2$
	1,0	$p(1-p)^2$
	1,1	$p(1-p)$
	1,2	$p^2(1-p)$
	2,1	$p^2(1-p)$
	2,2	p^3
2	0,0	$(1-p)^2$
	1,1	$2p(1-p)$
	2,2	p^2

$$E(X_1X_2) = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{k=0}^{2} ij\pi_k \Pr(X_1 = i, X_2 = j \mid IBD = k)$$
$$= (\pi_1 + 2\pi_2)p(1-p) + 4p^2.$$

It follows that $\text{cov}(X_1, X_2) = (\pi_1 + 2\pi_2)p(1-p) = 4\phi p(1-p)$.

We now derive the variance of the estimated allele frequency for data on a single family, ignoring the relationships between individuals. Since $\hat{p}_i^{(2)} = \Sigma_{j=1}^{k_i} X_{ij}/(2k_i)$, we have

$$\text{var}[\hat{p}_i^{(2)}] = \text{var}\left(\frac{\sum_j X_{ij}}{2k_i}\right)$$
$$= \frac{1}{4k_i^2}\left[\sum_j \text{var}(X_{ij}) + \sum_{j\neq j'} \text{cov}(X_{ij}, X_{ij'})\right]$$
$$= \frac{p(1-p)}{4k_i^2}\left(2k_i + 4\sum_{j\neq j'} \phi_{ijj'}\right)$$
$$= \bar{\phi}_i p(1-p).$$

Note that this formula applies to any set of individuals (not just siblings).

Genetic Epidemiology 23:465–466 (2002)

Erratum: Broman KW. 2001. Estimation of Allele Frequencies With Data on Sibships. Genet Epidemiol 20:307–15.

Karl W. Broman*

Department of Biostatistics, Johns Hopkins University, Baltimore Maryland

In the April 2001 issue of *Genetic Epidemiology*, in the article "Estimation of Allele Frequencies With Data on Sibships," by Broman (20:307–15), there is an error on page 310, in the second paragraph under "Method 3: Accounting for Relationships." The stated probabilities that an allele in the second sibling is not identical by descent (IBD) with one of the first sibling's alleles, written as $p_l/(1+p_l)$, are incorrect; we had missed two important cases. Let (g_{11}, g_{12}) denote the two alleles of the genotype of the first sibling, (g_{21}, g_{22}) denote the two alleles of the genotype of the second sibling, and $\boldsymbol{g} = (g_{11}, g_{12}, g_{21}, g_{22})$. Further, let $\boldsymbol{pg}$ denote the genotypes for the two parents, and A denote the event "g_{21} is not IBD with g_{11} or g_{12}." We seek $\Pr(A|\boldsymbol{g})$, which we calculate by conditioning on the parents' genotypes, as follows:

$$\Pr(A|\boldsymbol{g}) = \frac{\sum_{\boldsymbol{pg}} \Pr(\boldsymbol{pg}) \Pr(\boldsymbol{g}|\boldsymbol{pg}) \Pr(A|\boldsymbol{g}, \boldsymbol{pg})}{\sum_{\boldsymbol{pg}} \Pr(\boldsymbol{pg}) \Pr(\boldsymbol{g}|\boldsymbol{pg})}$$

The correct probabilities appear in Table I.

As a result of this error, the numbers in Table I of Broman (2001) were slightly wrong. We have rerun our computer simulations, after modifying our algorithm using the corrected probabilities, to obtain the true maximum likelihood estimates (MLEs). The corrected version of the table appears in Table II.

*Correspondence to: Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore MD 21205. E-mail: kbroman@jhsph.edu
Received for publication 10 July 2002; Revision accepted 11 July 2002
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/gepi.10194

TABLE I. Probability That an Allele in the Second Sibling Is Not IBD With Either Allele in the First Sibling

Sibs' genotypes			
g_{11} g_{12}	g_{21} g_{22}	Pr(g_{21} not IBD with g_{11} or $g_{21}\|g$)	Pr(g_{22} not IBD with g_{11} or $g_{21}\|g$)
11	11	$p_1/(1+p_1)$	$p_1/(1+p_1)$
11	22	1	1
11	12	$p_1/(1+p_1)$	1
11	23	1	1
12	12	$(p_1+2p_1p_2)/(1+p_1+p_2+2p_1p_2)$	$(p_2+2p_1p_2)/(1+p_1+p_2+2p_1p_2)$
12	13	$2p_1/(1+2p_1)$	1
12	34	1	1

TABLE II. Corrected Version of Table I in Broman (2001)

	p			
Het	0.05	0.10	0.15	0.20
0.70	1.46	1.44	1.42	1.39
0.75	1.47	1.45	1.42	1.39
0.80	1.47	1.43	1.42	1.40
0.85	1.48	1.43	1.42	1.40
0.90	1.48	1.46	1.43	1.42

It is interesting to note that the numbers in the corrected table are somewhat smaller than those in the original version. It appears that the true MLEs have somewhat greater standard deviations (SDs) than our flawed algorithm, especially for the larger values of the allele frequency, p. While our original estimates exhibit a slight negative bias (-0.001 in the case of $p = 0.2$ with heterozygosity $= 0.9$), they have smaller SDs and thus smaller mean square errors than the true MLEs. Note that the true MLEs appear to be unbiased.

We thank Mary Sara McPeek for identifying this error.