Sib-parentage testing using molecular markers when parents are unknown

D. García*, C. Carleos†, D. Parra* and J. Cañón*

*Laboratorio de Genética, Facultad de Veterinaria, Universidad Complutense de Madrid, Madrid, Spain. †Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Oviedo, Oviedo, Spain

Summary

The formulae for computing the so-called *Sib Index* using codominant alleles for (1) full-sib and (2) half-sib parentage are given. Hypothesis testing is based on the distribution of conditional likelihood ratio or Bayes' factor. Thresholds for rejecting the null hypothesis and *P*-values were obtained in function of the number of alleles and their frequency distributions. Simulations showed that a relatively low number of marker systems (e.g. 20) are enough to accept the hypothesis of sib parentage with a reasonable power for usual significance levels, but that a higher number would be necessary if full-sib against half-sib parentage is the contrast to be carried out. The effect of sampling variation on the allele frequencies on power calculations is also analysed.

Keywords Bayes' factor, molecular markers, sib parentage test.

Introduction

There is a need to check the correct paternity relationships to be used for predicting the genetic merit of the individuals included in the breeding programme through the numerator relationship matrix (Henderson 1976), especially for farmed animal species. It is well known that errors in paternity assignment delay genetic progress, whose magnitude under certain circumstances reaches that of the amount of paternity errors in the pedigree (Ron *et al*. 1996). While in these farmed species paternity testing is generally concerned with the exclusion of paternity, in others the requirement is to establish family relationships for legal, social or medical reasons (Pena & Chakraborty 1994). Exclusions of paternity are irrefutable and the power of a set of genetic markers to exclude is systematically computed into the exclusion probability, which depends on allelic frequencies in the population (Jamieson 1994; Jamieson & Taylor 1997). The availability for most domestic and many wild animal species of a large number of highly informative DNA markers and their utility for checking parentage have increased the possibilities (Goodnight & Queller 1999; Ritland 2000; Fiumera & Asmussen 2001). In particular, showing whether two individuals are sibs when no parental information is accessible is one of the most frequent questions asked for at animal genetic services or forensic laboratories. It is clear that under the situation, in which exclusion of paternity is the goal, the acceptance of the exclusion is irrefutable because, assuming no mutations and no laboratory errors, a descendent must carry parental alleles. However, proof of sib-ships depends on statistical inference. This paper presents formulae to compute *Sib* and *Half-sib Indexes* using codominant markers when parent information is unknown and their distributions under different amounts of marker information.

Address for correspondence

Javier Cañón, Laboratorio de Genética,
Facultad de Veterinaria, Universidad Complutense de Madrid,
28040 Madrid, Spain.
E-mail: jcanon@vet.ucm.es

Accepted for publication 29 March 2002

The conditional likelihood ratio

The conditional likelihood ratio (LR) or Bayes factor has traditionally been considered as the way to evaluate the evidence in paternity disputes (Aitken 1997). It is the ratio of two probabilities, the probability of G (the genotype of the individuals) when S (both individuals are sib related) is true [$\Pr(G/S)$] and the probability of G when $\bar{S}$ is false [$\Pr(G/\bar{S})$]. The likelihood ratio takes values between 0 and ∞, while its logarithm, which receives the name of *the weight of the evidence* (Good 1950), takes values on $(-\infty, \infty)$. It has been

argued in favour of the use of the logarithm of the likelihood ratio that the results it produces are in accordance with the weighing of evidence in the scale of justice.

The hypotheses that we wish to test are: (a) **S**, both individuals are sibs, (b) **I**, the individuals are unrelated random individuals from the referenced population, (c) **HS**, both individuals are half-sibs. Individuals are genotyped and the result is G. Large LR values argue in favour of sib-ships, whereas small values argue against it.

In the context of this paper the probability ratio is called *Sib Index* or *Half-sib Index*.

Calculation of the *Sib Index* for full-sib testing

Formally, our interest is to calculate the joint probability of two genotypes conditioned to the fact that the individuals carrying them are sibs, and compare it with the probability of those genotypes when the individuals are not sibs. In order to do that, and assuming that we are dealing with several loci, we will show first how to do it with a particular locus.

For a given locus A, let A_iA_j and A_kA_l be the genotypes for A of the two individuals, and denote by s the event of them being indeed whole sibs.

We can define the *Sib Index* for locus A as

$$SI_A = \frac{P\lfloor A_iA_j \wedge A_kA_l | s \rfloor}{P[A_iA_j \wedge A_kA_l | \neg s]}.$$

The probability in the denominator equals, under Hardy–Weinberg equilibrium $(2 - \delta_{ij})(2 - \delta_{kl})p_ip_jp_kp_l$, where p_m stands for the frequency of allele m in the population, $m = 1,\ldots,n$, n being the number of alleles in locus A, and δ_{mn} is the well-known Kronecker Delta, which equals 1 if $m = n$ and 0 otherwise. This result is obvious because both individuals are independent in this situation.

To obtain the probability in the numerator, we need a little algebra. We know that $P[A_iA_j \wedge A_kA_l|s] = P[A_kA_l|s \wedge A_iA_j]P[A_iA_j|s]$. Now, s gives no information for the calculus of $P[A_iA_j]$, hence $P[A_iA_j|s] = (2 - \delta_{ij})p_ip_j$ – remember we are assuming HW equilibrium.

We are therefore interested in obtaining the probability of the genotype of one sib conditioned to the fact that it is indeed a sib of the other and to the genotype of the latter. Now,

$$P[A_kA_l|s \wedge A_iA_j] = \sum_{\alpha,\beta,\gamma,\delta=1}^{n} P[A_kA_l \wedge GP = (A_\alpha A_\beta \times A_\gamma A_\delta)|s \wedge A_iA_j],$$

where GP is the genotype of a hypothetical parental mating. But

$$\sum_{\alpha,\beta,\gamma,\delta=1}^{n} P\left[A_kA_l \wedge GP = (A_\alpha A_\beta \times A_\gamma A_\delta)|s \wedge A_iA_j\right]$$
$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{n} P\left[A_kA_l|GP = (A_\alpha A_\beta \times A_\gamma A_\delta) \wedge s \wedge A_iA_j\right]$$
$$P\left[GP = (A_\alpha A_\beta \times A_\gamma A_\delta)|s \wedge A_iA_j\right].$$

The sum is taken through all the possible matings for a locus with n alleles, but, as $P\lfloor GP = (A_\alpha A_\beta \times A_\gamma A_\delta)|s \wedge A_iA_j \rfloor \neq 0$ only for those matings in which one parent carries at least one A_i allele and the other carries at least one A_j allele, it reduces to

$$\sum_{\alpha,\beta=1}^{n} P\left[A_kA_l|GP = (A_iA_\alpha \times A_\beta A_j) \wedge s \wedge A_iA_j\right]$$
$$P\left[GP = (A_iA_\alpha \times A_\beta A_j)|s \wedge A_iA_j\right].$$

Now, once the alleles A_i and A_j are fixed in the genotypes of the parents, the only stochastic variation is due to A_α and A_β, so

$$P\lfloor GP = (A_iA_\alpha \times A_\beta A_j)|s \wedge A_iA_j \rfloor = p_\alpha p_\beta,$$

and therefore

$$P[A_kA_l|s \wedge A_iA_j]$$
$$= \sum_{\alpha,\beta=1}^{n} p_\alpha p_\beta P[A_kA_l|GP = (A_iA_\alpha \times A_\beta A_j) \wedge s \wedge A_iA_j].$$

For example, consider a locus A with three alleles A_1, A_2 and A_3, and suppose that the reference individual is A_1A_2, his putative sib A_1A_3 and the alleles have frequencies of p_1, p_2 and p_3. Then the possible matings to originate an A_1A_2 individual are

$$\begin{bmatrix} A_1A_1 \\ A_1A_2 \\ A_1A_3 \end{bmatrix} \times \begin{bmatrix} A_1A_2 \\ A_2A_2 \\ A_3A_2 \end{bmatrix}$$

and thus

$$\begin{aligned} P[A_1A_3|s \wedge A_1A_2] &= 0.5p_1p_3 + 0.25p_3p_2 + 0.25p_3^2 \\ &\quad + 0.25p_1p_3 \\ &= 0.5p_1p_3 + 0.25p_3(p_1 + p_2 + p_3) \\ &= 0.5p_1p_3 + 0.25p_3. \end{aligned}$$

It can be easily shown that, in general,

$$\begin{aligned} P(A_kA_l|s \wedge A_iA_j) &= 0.25((2 - \delta_{kl})p_kp_l + \delta_{ik}(1 + \delta_{ij})p_l \\ &\quad + \delta_{jl}(1 - \delta_{ij})p_k + \delta_{ik}\delta_{jl}) \end{aligned}$$

$\forall\ i, j, k, l \in \{1,\ldots, n\}$, n being the number of alleles in the locus.

This formula was obtained by trying to provide a unique expression for the explicit, separate formulae deduced for

Table 1 The nine possible cases for $P(A_kA_l|s \wedge A_iA_j)$. In rows the possibilities for the individual in the condition; in columns the possibilities for the individual whose genotype's probability is calculated.

	Homozygous sharing two alleles (A_iA_i)	Homozygous sharing one allele (A_iA_i)	Homozygous sharing zero alleles (A_kA_k)	Heterozygous sharing two alleles (A_iA_j)	Heterozygous sharing one allele (A_iA_l)	Heterozygous sharing zero alleles (A_kA_l)
Homozygous (A_iA_i)	$0.25(1 + p_i)^2$		$0.25p_k^2$		$0.5(p_ip_l + p_l)$	$0.5p_kp_l$
Heterozygous (A_iA_j)		$0.25p_i(1 + p_i)$	$0.25p_k^2$	$0.5[p_ip_j + 0.5(1 + p_i + p_j)]$	$0.5(p_ip_l + 0.5p_l)$	$0.5p_kp_l$

each sibs' genotype combination, depicted in Table 1, for computer programming purposes. It is important to note that, in case of being implemented in a programme, the genotypes to be the input in the formula have to be stored in the right way, as denoted in Table 1. For example, to compute, $P(A_jA_i|s \wedge A_iA_j)$, the haplotype A_jA_i has to be input as A_iA_j. Otherwise the formula will not work properly. Of course it is possible to generalize a step beyond, but the expression thereby obtained is far more obtuse.

Note that, as expected, the sum of the probabilities along the rows equals 1.

The *Sib Index* for locus A would thus be

$$SI_A = \frac{(2 - \delta_{kl})p_kp_l + \delta_{ik}(1 + \delta_{ij})p_l + \delta_{jl}(1 - \delta_{ij})p_k + \delta_{ik}\delta_{jl}}{4(2 - \delta_{kl})p_kp_l}.$$

When considering multiple independent loci, as usually happens in the situations in which this procedures are useful, the joint *Sib Index* would be the product of the different indexes. Hence, $SI = \prod_{m=1}^{M} SI_m$.

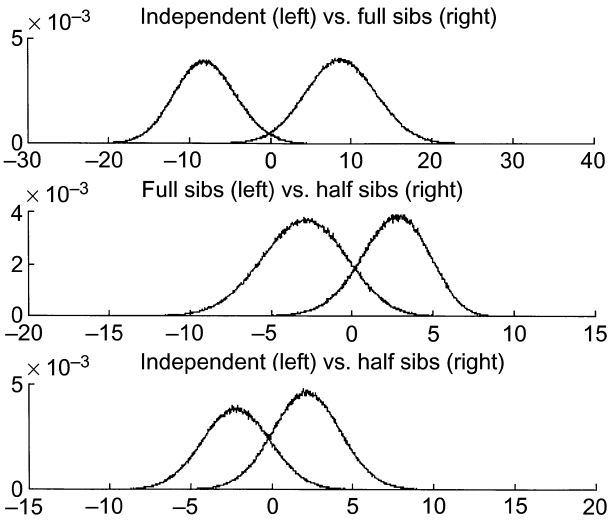

Figure 1 Empirical probability density functions of the log-Sib Index (up) and log-Half-sib Index (down) under the two hypotheses tested in each (null hypothesis on left). In the middle, empirical pdf of the Log-PR of being full sibs (H_0) vs. being half-sibs. All of them obtained for a random set of frequencies for 20 markers with five alleles each and 10^6 simulations.

Note that although this rationale is based on the calculation of a conditional probability on the genotype of one of the putative sibs in order to obtain the joint probability of the two putative sibs' genotype, the choice of the genotype upon which the probability is to be conditioned has no influence on the final result.

Calculation of the *Sib Index* for half-sib testing

Although with some variations that we will discuss later on, the situation is fairly analogous to the full-sib case. Assuming the same locus notation as in the previous section, we look for the value of the quotient $\frac{P[A_iA_j \wedge A_kA_l|hs]}{P[A_iA_j \wedge A_kA_l|\neg hs]}$, which will be called the Half-sib Index for locus A, and denoted as HSI_A, and where hs denotes the event of the two individuals being half-sibs. For the same reasons as above, we can restrict ourselves to calculate $p\lfloor A_kA_l|hs \wedge A_iA_j\rfloor$. Now, instead of expanding this probability by intersecting with the genotypes of the parents, we will do it only with the genotype of the father, denoted as GF, because both individuals share the same father, but not necessarily the same mother. Therefore,

$$P[A_kA_l|hs \wedge A_iA_j] = \sum_{\alpha,\beta=1}^{n} P\left[A_kA_l \wedge GF = A_\alpha A_\beta|hs \wedge A_iA_j\right].$$

Now,

$$\begin{aligned}
&\sum_{\substack{\alpha=1\\ \beta\geq\alpha}}^{n} P\left[A_kA_l \wedge GF = A_\alpha A_\beta|hs \wedge A_iA_j\right] \\
&= \sum_{\substack{\alpha=1\\ \beta\geq\alpha}}^{n} P\left[A_kA_l|GF = A_\alpha A_\beta \wedge hs \wedge A_iA_j\right] P\left[GF = A_\alpha A_\beta|hs \wedge A_iA_j\right] \\
&= \sum_{\alpha=1}^{n} [P[A_kA_l|GF = A_iA_\alpha \wedge hs \wedge A_iA_j] P[GF = A_jA_\alpha|hs \wedge A_iA_j] \\
&\quad + (1 - \delta_{ij}) P\lfloor A_kA_l|GF = A_jA_\alpha \wedge hs \wedge A_iA_j\rfloor \\
&\quad P\lfloor GF = A_jA_\alpha|hs \wedge A_iA_j\rfloor] \\
&= \sum_{\alpha=1}^{n} 0.5(1 + \delta_{ij})p_\alpha [P[A_kA_l|GF = A_iA_\alpha \wedge hs \wedge A_iA_j] \\
&\quad + (1 - \delta_{ij}) P\lfloor A_kA_l|GF = A_jA_\alpha \wedge hs \wedge A_iA_j.\rfloor]
\end{aligned}$$

Consider the simplest example, in which one individual bears genotype A_1A_2, and the other A_3A_4, all the alleles

being different from each other, we try to find the probability of genotype A_3A_4 conditioned to the fact that a half-sib of him carries A_1A_2. Then

$$P[A_3A_4|hs \wedge A_1A_2] = 0.5p_3[P[A_3A_4|GF = A_1A_3] + P[A_3A_4|GF = A_2A_3]] + 0.5p_4[P[A_3A_4|GF = A_1A_4] + P[A_3A_4|GF = A_2A_4]].$$

Now, $P[A_3A_4\,|\,GF = A_1A_3]$ is the probability of the individual carrying allele A_4 and the father passing A_3, which equals $1/2p_4$. Therefore, $P[A_3A_4|hs \wedge A_1A_2] = 0.5p_3[1/2p_4 + 1/2p_4] + 1/2p_4[1/2p_3 + 1/2p_3] = p_3p_4$.

As in the previous situation, we can build up a table with all the possible cases, such as Table 2.

As before, it is easy to show that the probabilities in each row sum one, as expected from a well defined probability measure.

Again, a more general formula can be induced from the table:

$$P(A_kA_l|hs \wedge A_iA_j) = 0.25(2(2 - \delta_{kl})p_kp_l + \delta_{ik}(1 + \delta_{ij})p_l + \delta_{jl}(1 - \delta_{ij})p_k),$$

and the same comments of the previous case regarding alleles ordering in the haplotypes apply here.

Thus, the Half-sib Index for locus A is defined as

$$HSI_A = \frac{2(2 - \delta_{kl})p_kp_l + \delta_{ik}(1 + \delta_{ij})p_l + \delta_{jl}(1 - \delta_{ij})p_k}{4(2 - \delta_{kl})p_kp_l},$$

and the joint *Half-sib Index* for multiple and independent loci is

$$HSI = \prod_{m=1}^{M} HSI_m.$$

Hypothesis testing

Simulation routines were developed to study the performance of the test under several situations. Empirical power values and probability density functions were obtained for different marker configurations, allele frequency situations and significance levels. Allele frequencies were simulated for populations with 10, 20 and 30 markers, all of them having the same number of alleles, either 5 or 10. The sets of frequencies were generated in three different ways: at random, uniformly with slight variations and uniformly with slight variations except but some extreme frequency alleles. Uniform with slight variations means here that if the number of alleles for a certain locus is n, each allele is present not with frequency $1/n$, but $1/n \pm 0.01$ ($f_1 = 1/n + 0.01$, $f_2 = 1/n - 0.01$, and so on, with $f_n = 1/n - 0.01$ if n is even and $f_n = 1/n$ if n is odd). Uniform with slight variations but with some extreme frequency alleles means the same as above except that 20% of the loci involved are considered to have extreme frequency alleles, that is, all the alleles but one are present with frequency 0.01. The power was obtained for three significance levels: 0.001, 0.0001 and 0.00001. To obtain the power of one test for a set of allele frequencies, 100 000 pairs of individuals were simulated with those frequencies under null and another 100 000 under alternate hypotheses. Formulae in Table 1 or 2, depending on the contrast, were applied to the 100 000 pairs on each hypothesis, and thus 200 000 values of the statistic were obtained, 100 000 under each hypothesis. For a significance level α, the $(1 - \alpha) \times 100\,000$th highest value of the 100 000 H_0 values was taken as the rejection threshold, and so the power was obtained as the proportion of H_1 values above that threshold. Probability density values were empirically approximated by taking histogram values as pdf values, with histogram values calculated for 1000 narrow intervals. The method was applied as well in a real Irish-Setter dog breed population, to test the parentage of two putative full sibs. Allelic frequencies were estimated from a total of 64 individuals.

Tables 3–5 show the figures for the different scenarios. As expected, the increase of information from markers results in an increase in power. Reasonable values are obtained for full-sib parentage determination for a non-rare situation of 20 markers with five alleles each. If the number of markers grows up to 30, maintaining the plausible value of five alleles per marker locus, it makes the test show good levels of power, even for a significance level as low as 10^{-4}. Raising the number of alleles per locus would keep the power high even for $\alpha = 10^{-5}$. As depicted in Fig. 2, in the perhaps overly optimistic situation of having 10 alleles per locus for a total of 30 markers, the density functions of the natural logarithm of

Table 2 The nine possible cases for $P(A_kA_l|hs \wedge A_iA_j)$. In rows the possibilities for the individual in the condition; in columns the possibilities for the individual whose genotype's probability is calculated.

	Homozygous sharing two alleles (A_iA_i)	Homozygous sharing one allele (A_iA_i)	Homozygous sharing zero alleles (A_kA_k)	Heterozygous sharing two alleles (A_iA_j)	Heterozygous sharing one allele (A_iA_l)	Heterozygous sharing zero alleles (A_kA_l)
Homozygous (A_iA_i)	$0.5p_i(1 + p_i)$		$0.5p_k^2$		$0.5(2p_ip_l + p_l)$	p_kp_l
Heterozygous (A_iA_j)		$0.25p_i(1 + 2p_i)$	$0.5p_k^2$	$0.5[2p_ip_j + 0.5\,(p_i + p_j)]$	$0.5(2p_ip_l + 0.5p_l)$	p_kp_l

Table 3 Power values for independent vs. full-sibs hypothesis testing.

	10 markers						20 markers						30 markers					
	5 al/mark			10 al/mark			5 al/mark			10 al/mark			5 al/mark			10 al/mark		
	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$
Random	0.44536	0.22934	0.10157	0.79854	0.61631	0.44069	0.85531	0.66939	0.47467	0.99143	0.96517	0.91936	0.97811	0.92002	0.82734	0.99972	0.99813	0.99286
Ps-uniform	0.53228	0.30110	0.14205	0.85603	0.69940	0.49415	0.91370	0.76140	0.57378	0.99584	0.98165	0.94990	0.99043	0.95893	0.90395	0.99992	0.99946	0.99766
Ps-unf w/ext	0.43149	0.21368	0.09171	0.77649	0.58916	0.41665	0.83809	0.65952	0.43453	0.98641	0.95508	0.88905	0.96812	0.89104	0.74534	0.99954	0.99777	0.98862

Table 4 Power values for independent vs. half-sibs hypothesis testing.

	10 markers						20 markers						30 markers					
	5 al/mark			10 al/mark			5 al/mark			10 al/mark			5 al/mark			10 al/mark		
	$\omega\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$
Random	0.05431	0.01266	0.00274	0.16089	0.05476	0.01901	0.15694	0.05362	0.01610	0.45477	0.23445	0.12645	0.29679	0.11638	0.03599	0.69237	0.44494	0.26397
Ps-uniform	0.05494	0.01323	0.00315	0.18322	0.05651	0.01614	0.17761	0.05811	0.01249	0.51110	0.25996	0.11961	0.33914	0.14846	0.06414	0.76778	0.53264	0.30665
Ps-unf w/ext	0.05051	0.01444	0.00328	0.15367	0.05228	0.02039	0.14791	0.05199	0.02231	0.43077	0.23121	0.10272	0.27401	0.11076	0.03512	0.68239	0.42432	0.23587

Table 5 Power values for full-sibs vs. half-sibs hypothesis testing.

	10 markers						20 markers						30 markers					
	5 al/mark			10 al/mark			5 al/mark			10 al/mark			5 al/mark			10 al/mark		
	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$
Random	0.05166	0.00818	0.01555	0.08469	0.01600	0.00267	0.18472	0.05909	0.01636	0.37472	0.12714	0.03166	0.38957	0.15407	0.04898	0.64822	0.36337	0.14694
Ps-uniform	0.06028	0.01067	0.00165	0.09356	0.02013	0.00549	0.24025	0.07720	0.01720	0.42962	0.15788	0.04077	0.47114	0.22348	0.09440	0.73880	0.43798	0.18055
Ps-unf w/ext	0.04208	0.00727	0.00146	0.06632	0.01209	0.00195	0.16765	0.05008	0.01352	0.30247	0.09316	0.03177	0.35183	0.14299	0.05101	0.59305	0.29907	0.08477

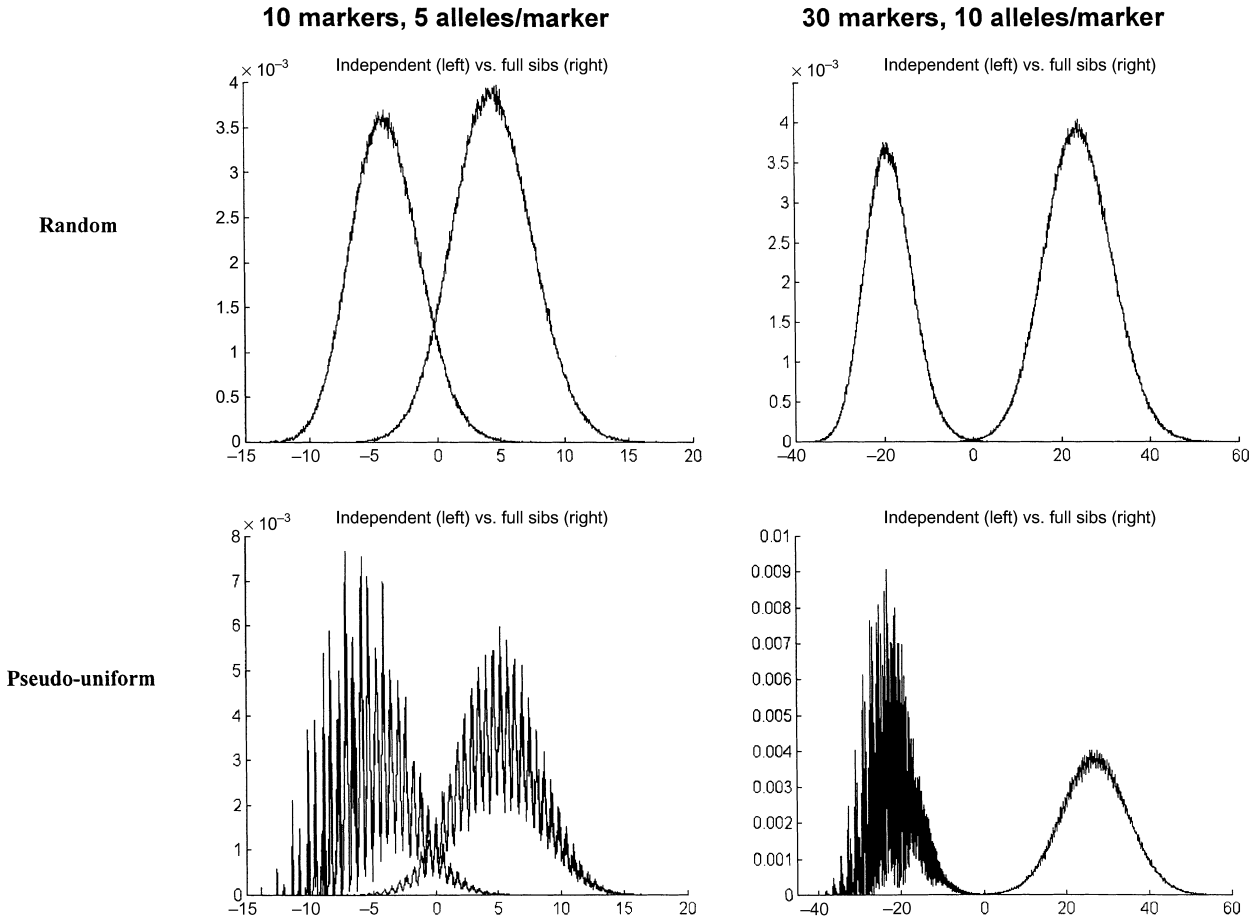

Figure 2 Empirical probability density functions of the log-Sib Index for extreme marker situations: the least informative (10 markers, five alleles/marker) in the left column and the most (30 markers, 10 alleles/marker) in the right. Pdfs were obtained for random sets of allelic frequencies (upper row) and pseudo-uniform ones (lower row).

the test under the null (independent individuals) and alternate (full-sibs) hypotheses have almost no intersection, which makes the contrast have very little error. Results also suggest that the test performs better for uniform allelic frequency distributions than for the cases in which alleles with extreme frequencies are involved.

On the other hand, power values for half-sib testage, both vs. independency and vs. full-sibship are not very encouraging, because for optimum conditions of having 30 markers with 10 alleles each and the lowest significance level, power gets to about 75% in the best cases. This outcome was expected, as testing for half-sibship with no parent information implies allowing for different genotypes on the non-shared parent, so there is a loss of information which lowers the power of the test. Figure 2 shows this situation more graphically, because distributions under null and alternate hypothesis share much of their support. In any case, for individual purposes, testing is always advisable, as it may happen that the individuals are in fact genetically distant and *P*-values result conclusive.

Figure 3 shows the plots for a real data set from an Irish-Setter dog breed population. A total of 18 molecular markers (microsatellites) were analysed, with the number of alleles ranging from 3 to 14, the mean number being 6.94 and a mean expected heterozygosity value of 0.64. Table 6 shows the figures for power in this situation. Reasonable power is obtained only for full-sibship vs. independency comparisons for $\alpha = 0.001$.

Simulations were also performed to determine whether sampling variation on the allele frequencies affect power calculations. First, 1000 sets of random allele frequencies for 15 markers, each of them with seven alleles, were generated, and for each set, the power for the independent vs. full sibs test was calculated. The mean power among the 1000 sets was of 0.864 ± 0.010 for $\alpha = 0.001$. This shows that very different sets of allele frequencies provide similar power estimates, so the power of the test relies in the

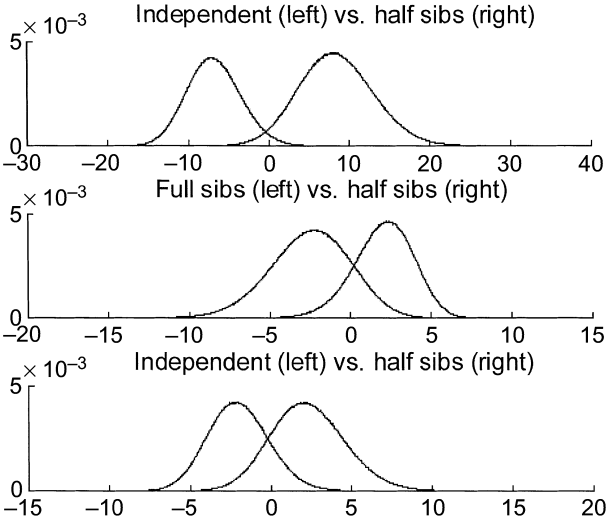

Figure 3 Same as Figure 1 for real data from an Irish Setter dog breed population.

Table 6 Power values for real Setter dog breed population data.

	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$
Independent vs. full-sibs	0.8155	0.6269	0.4532
Independent vs. half-sibs	0.1879	0.0707	0.0233
Full sibs vs. half-sibs	0.1167	0.0290	0.0069

number of markers and alleles, more than in the frequencies distribution. For the Irish-Setter population, the estimated frequencies were taken as true values and from them 1000 samples of 25, 50 and 100 individuals, respectively, were generated. For each sample of individuals allelic frequencies were re-estimated, and the new estimations were used to calculate the power of the independent vs. full sibs test. The mean power for $\alpha = 0.001$ was of 0.775 ± 0.017 for a 25-individual sampling scheme, 0.796 ± 0.014 for a 50-individual one and 0.806 ± 0.012 for the 100 samples. We see then that multiplying by four the sample size to estimate the allelic frequencies increases the accuracy of power estimation by only 0.5%, which supports our previous conclusion. We can also conclude that, as expected, the higher the sample size, the more accurate the estimation in terms of deviation with respect to the assumed *real* value of 0.816 (see Table 6). Nevertheless, with a reasonably low sample size of 50 individuals the mean error in the power estimation is only 2.5%.

Other authors have developed similar methods. Goodnight & Queller (1999) presented *Kinship*, a software for performing likelihood tests of pedigree relationships. Their method, however, is based on inaccurate probability calculations. Take, for instance, the most unquestionably wrong case, that of two supposed unrelated individuals carrying four different alleles for a certain marker, say A_1A_2 and A_3A_4. It is clear that the likelihood of the two genotypes under the non-relationship hypothesis must be $2p_1p_2 \times 2p_3p_4 = 4p_1p_2p_3p_4$, where p_i stands for the frequency of A_i, $i = 1,\ldots,4$. According to Goodnight and Queller's procedure and the formulae in Table 2 of Goodnight & Queller (1999), the likelihood would equal $\frac{4}{4}\ p_1p_2p_3p_4 = p_1p_2p_3p_4$, which is clearly incorrect. Probably the inaccuracy lies in the *allele by allele* approach that they followed in the deduction of their formulae. In any case, their results are not comparable with ours, because they calculate the number of loci needed for a power of 0.5 and a significance level of 0.05, which is too high a figure for the usual forensic applications. Furthermore, they consider loci with 20 equally frequent alleles per locus, which is a certainly unrealistic situation to simulate.

Conclusions

Formulae are given to compute the probabilities ratio for different hypotheses when sibs or half-sibs are implied and parents are not known. Results, obtained under the Hardy–Weinberg assumptions and assuming that population allele frequencies are known without error, show that the amount of information generally used by the service laboratories can be sufficient to test full-sib or, with some less certainty, half-sib parentage, as more marker information will be required to reach equivalent power.

Acknowledgements

This work received the financial support of the EC DGVI QLRT-99-30147. Blood samples from the Irish-Setter breed were provided by the Federación Española de Caza.

References

Aitken C.G.G. (1997) *Statistics and the Evaluation of the Evidence for Forensic Scientist*. John Wiley & Sons, New York.

Fiumera A.C. & Asmussen M.A. (2001) Difficulties in parentage analysis: the probability that an offspring and parent have the same heterozygous genotype. *Genetical Research, Canberra* **78,** 163–70.

Good I.J. (1950) *Probability and the Weighing of Evidence*, Charles Griffing Limited, London.

Goodnight K.F. & Queller D.C. (1999) Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology* **8,** 1231–4.

Henderson C.R. (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32,** 69–83.

Jamieson A. (1994) The effectiveness of using co-dominant polymorphic allelic series for (1) checking pedigrees and (2) distinguishing full-sib pair members. *Animal Genetics* **25,** 37–44.

Jamieson A. & Taylor St.C.S. (1997) Comparison of three probability formulae for parentage exclusion. *Animal Genetics* **28,** 397–400.

Pena S.D.J. & Chakraborty R. (1994) Paternity testing in the DNA era. *Trends in Genetics* **10,** 204–9.

Ritland K. (2000) Marker-inferred relatedness as a toll for detecting heritability in nature. *Molecular Ecology* **9,** 1195–204.

Ron M., Blanc Y., Band M., Ezra E. & Weller J.I. (1996) Misidentification rate in the Israeli dairy cattle population and its implications for genetic improvement. *Journal of Dairy Science* **79,** 676–81.