Genet. Res., Camb. (2001), **78**, *pp.* 163–170. *With* 4 *figures.* © 2001 Cambridge University Press
DOI: 10.1017/S0016672301005237 Printed in the United Kingdom

Difficulties in parentage analysis: the probability that an offspring and parent have the same heterozygous genotype

A. C. FIUMERA AND M. A. ASMUSSEN*
Department of Genetics, University of Georgia, Athens, GA 30602, USA

(*Received 2 March 2001 and in revised form 2 May 2001*)

Summary

Parentage studies often estimate the number of parents contributing to half-sib progeny arrays by counting the number of alleles attributed to unshared parents. This approach is compromised when an offspring has the same heterozygous genotype as the shared parent, for then the contribution of the unshared parent cannot be unambiguously deduced. To determine how often such cases occur, formulae for co-dominant markers with n alleles are derived here for P_h, the probability that a *given heterozygous parent* has an offspring with the same heterozygous genotype, and P_a, the probability that a *randomly chosen offspring* has the same heterozygous genotype as the shared parent. These formulae have been derived assuming Mendelian segregation with either (1) an arbitrary mating system, (2) random mating or (3) mixed mating. The maximum value of P_a under random mating is 0·25 and occurs with any two alleles each at a frequency of 0·5. The behaviour with partial selfing (where reproduction is by selfing with probability s, and random mating otherwise) is more complex. For $n \leqslant 3$ alleles, the maximum value of P_a occurs with any two alleles each at a frequency of 0·5 if $s < 0{\cdot}25$, and with three equally frequent alleles otherwise. Numerically, the maximum value of P_a for $n \geqslant 4$ alleles occurs with $n^* \leqslant n$ alleles at equal frequencies, where the maximizing number of alleles n^* is an increasing function of the selfing rate. Analytically, the maximum occurs with all n alleles present and equally frequent if $s \geqslant 2/3$. In addition, the potential applicability of these formulae for evolutionary studies is briefly discussed.

1. Introduction

Empirical studies of mating systems often attempt to estimate the true number of parents contributing to half-sib progeny arrays (Levine *et al.*, 1980; Parker & Kornfield, 1996; Jones & Avise, 1997; Moran & Garcia-Vazquez, 1998; Bollmer *et al.*, 1999). This situation frequently arises in both plants and animals when a set of offspring share one parent in common but could have multiple unshared parents. For example, a litter in mice would have the mother as the shared parent, but different offspring might have been sired by different fathers (Baker *et al.*, 1999). Several methods exist to estimate the true number of unshared parents contributing to such half-sib arrays (Levine *et al.*, 1980; Parker & Kornfield, 1996; Kellogg *et al.*, 1998; DeWoody *et al.*, 2000). For many of these methods, the estimate is derived from the number of distinct gametotypes (single locus alleles or multilocus haplotypes) that were contributed by the unshared parents; this is deduced by subtracting out the allele that was contributed by the shared parent. For instance, if at a given locus the shared parent of a half-sib progeny array has the genotype A_iA_j and an offspring has the genotype A_iA_k, then the shared parent contributed the A_i allele and, thus, the unshared parent contributed the A_k allele.

Difficulties can arise, however, when attempting to determine the contribution of the unshared parent. As observed by many researchers (Kellogg *et al.*, 1998; Kichler *et al.*, 1999; Fiumera *et al.*, 2001), if the shared parent and one of its offspring both have the same heterozygous genotype, then the contribution of the shared parent and therefore that of the unshared parent cannot be unambiguously determined. For example, if both the shared parent and one of its offspring have the diploid genotype A_iA_j, then it

* Corresponding author. Tel: +1 (706) 542 1455. Fax: +1 (706) 542 3910. e-mail: asmussen@arches.uga.edu

cannot be deduced whether the shared parent contributed the A_i or the A_j allele. If these ambiguous cases are common, they will hinder attempts to estimate the true number of parents contributing to half-sib progeny arrays based on counting the number of gametotypes contributed by the unshared parents.

Given this potential difficulty, it would be useful to have simple analytic formulae for the probability that a given heterozygous shared parent of genotype A_iA_j has an offspring of the same heterozygous genotype A_iA_j. In addition, it would be valuable to be able to calculate easily the probability that a randomly chosen offspring from a randomly chosen progeny array has the same heterozygous genotype as the shared parent. Researchers have calculated the first probability in the process of deriving exclusion probabilities for specific allelic systems (Boyd, 1955; Weiner, 1968) and assuming random mating (Evett & Weir, 1998). To our knowledge, no published formulae exist for the second value, and no formulae exist for either of these probabilities for the general case of co-dominant markers with n alleles and an arbitrary mating system.

Here we derive explicit analytic formulae for the n-allele case for (1) the probability, P_h, that a *given heterozygous parent* has an offspring with the same heterozygous genotype and (2) the overall probability of ambiguity, P_a, which is the probability that a *randomly chosen offspring* has the same heterozygous genotype as the shared parent. We first derive the general formulae assuming only Mendelian segregation and then present simplifications for the special cases of fully random mating and mixed-mating populations. For ease of exposition, we will refer to the shared parent as the female (♀) and the unshared parent as the male (♂), but analogous derivations and formulae apply if the male is the shared parent.

2. The probability of ambiguous offspring

(i) *General case with arbitrary mating system*

We first derive the probability, $P_h = P(O = A_iA_j|♀ = A_iA_j)$, that a heterozygous parent (♀) of genotype A_iA_j has an offspring (O) of the same heterozygous genotype A_iA_j for the general case with an arbitrary mating system. Assume Mendelian segregation at a single, autosomal, diploid locus with n alleles (A_1, A_2, …, A_n) where the frequency of allele A_i is p_i for $i = 1, 2, \ldots, n$ and $p_1 + p_2 + \cdots + p_n = 1$. Conditioning on all possible types of male genotypes and substituting in the expected frequency of A_iA_j offspring for each of the respective matings shows that

$$\begin{aligned} P_h &= \sum_{k \leqslant m} P(O = A_iA_j|A_iA_j♀ \times A_kA_m♂) \\ &\quad \times P(♂ = A_kA_m|♀ = A_iA_j) \\ &= \tfrac{1}{2}P(♂ = A_iA_i|♀ = A_iA_j) + \tfrac{1}{2}P(♂ = A_jA_j|♀ = A_iA_j) \\ &\quad + \tfrac{1}{2}P(♂ = A_iA_j|♀ = A_iA_j) \\ &\quad + \sum_{k \neq i,j} \tfrac{1}{4}p(♂ = A_iA_k|♀ = A_iA_j) \\ &\quad + \sum_{k \neq i,j} \tfrac{1}{4}P(♂ = A_jA_k|♀ = A_iA_j). \end{aligned} \quad (1)$$

The second quantity of interest, the probability, $P_a = \sum_{i<j} P(O = ♀ = A_iA_j)$, that a randomly chosen offspring has the same heterozygous genotype as the shared parent, is then obtained by conditioning on the genotype of the female,

$$P_a = \sum_{i<j} P(O = A_iA_j|♀ = A_iA_j)P(♀ = A_iA_j), \quad (2)$$

where $P(O = A_iA_j|♀ = A_iA_j) = P_h$ is given by equation (1) and $P(♀ = A_iA_j)$ is the female genotype frequency at this locus.

(ii) *Random mating*

Equations (1) and (2) take on a simple form when there is random mating with respect to genotype at this locus and Hardy–Weinberg frequencies. For this case, P_h can be derived by substituting the latter into equation (1), or simply by noting that an A_iA_j female can either pass on her A_i allele (with probability 0·5) and the male the A_j allele (with probability p_j), or her A_j allele (with probability 0·5) and the male the A_i allele (with probability p_i). Thus, for a random mating population at Hardy–Weinberg equilibrium, the probability that a given heterozygous female (A_iA_j) has an offspring with the same heterozygous genotype (A_iA_j) is

$$P_h = \tfrac{1}{2}(p_i + p_j), \quad (3)$$

where p_i and p_j are the frequencies of the alleles A_i and A_j, respectively. This formula can also be derived via descent measures (Evett & Weir, 1998, p. 117). Substituting equation (3) and the Hardy–Weinberg genotype frequencies into equation (2) (see Appendix A), we then find that the corresponding probability that a randomly chosen offspring has the same heterozygous genotype as the shared parent is

$$P_a = \sum_{i=1}^{n} p_i^2(1 - p_i). \quad (4)$$

(iii) *Mixed mating*

Equations (1) and (2) are also substantially simplified in mixed-mating populations where a proportion of females s self-fertilize and the remaining females $(1-s)$ outcross, where $0 < s < 1$. P_h here follows directly from the law of total probability (Ross, 1997) by conditioning on whether the female selfs or outcrosses. The probability that a given heterozygous

https://doi.org/10.1017/S0016672301005237 Published online by Cambridge University Press

female has an offspring of the same heterozygous genotype is thus

$$P_h = \tfrac{1}{2}[s+(1-s)(p_i+p_j)], \quad (5)$$

which is simply the weighted average of the values under pure selfing (1/2) and random mating, $\frac{1}{2}(p_i+p_j)$.

Substituting equation (5) and the mixed-mating equilibrium genotype frequencies of heterozygotes (Marshall & Weir, 1979) into equation (2) then reveals that the probability that a randomly chosen offspring has the same heterozygous genotype as the shared parent is

$$P_a = \left(\frac{1-s}{2-s}\right)\left[s\left(1-\sum_{i=1}^{n} p_i^2\right) + 2(1-s)\sum_{i=1}^{n} p_i^2(1-p_i) \right]. \quad (6)$$

If $s = 0$, equations (5) and (6) reduce to the random mating equations (3) and (4) respectively.

(iv) *Multiple loci*

Most empirical studies of parentage utilize data from multiple, unlinked molecular markers. It would be useful to calculate the probability, P_T, that a randomly chosen offspring will *not* have the same heterozygous genotype as the shared parent at any of the multiple loci surveyed. Under this condition, the contribution of the unshared parent will be unambiguous at all the loci. As the markers are independent, P_T is given by simply multiplying their probabilities $(1-P_a)$ together.

3. Factors affecting the probability of ambiguity

(i) *Effects of allele frequency distribution*

How frequently ambiguous offspring occur is determined by the maximum and minimum values of P_h and P_a. The probability that a given heterozygous female has an offspring of the same heterozygous genotype, P_h from equations (3) and (5), is a monotonically increasing function of the combined frequencies of the two associated alleles (p_i+p_j), ranging in both random and mixed-mating populations from a greatest lower bound of 0 for two rare alleles to a maximum of 1/2 when these are the only alleles in the population.

The behaviour of the overall probability of ambiguity, P_a, is more complex (Appendix B). We first consider a random mating population. The maximum of P_a is then 0·25 and occurs with two alleles, each at a frequency of 0·5. For any number of alleles n, P_a approaches this maximum as the allele frequency distribution approaches that of two equally frequent alleles with all other alleles at a frequency near 0 (see Fig. 1 and 2 for examples with two and three alleles). As expected, P_a has a minimum value of 0, which occurs when the locus is monomorphic (Figs. 1, 2). In addition, with equally frequent alleles, P_a rapidly decreases from 0·25 to 0 according to the formula $P_a = (1/n)(1-1/n)$, as the number of alleles (n) at the locus increases from two to infinity. Finally, as the number of marker loci with equally frequent alleles increases, P_T, the probability that a randomly chosen offspring does *not* have the same heterozygous genotype as the shared parent at any of the loci decreases (Fig. 3).

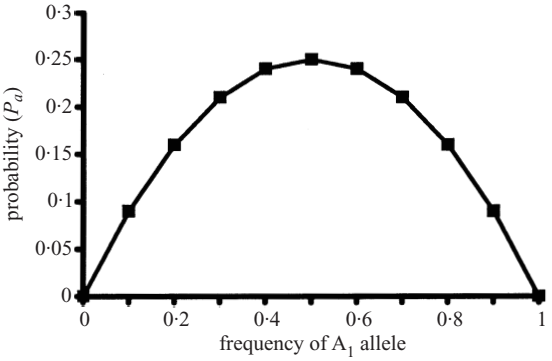

Fig. 1. Probability (P_a from equation (4)) that a randomly chosen offspring has the same heterozygous genotype as the shared parent in a random mating population, for $n = 2$ alleles. P_a is simply $p_i(1-p_i)$ which has a maximum value of 0·25 when both alleles have a frequency of 0·5, and declines to 0 as the frequency of the A_1 allele approaches either 0 or 1.

The situation is considerably more complex in mixed-mating populations (Fig. 4). Although the minimum value of P_a is still always zero and occurs when the locus is monomorphic or the selfing rate is 1 (because there are then no heterozygotes at equilibrium), the maximum value is governed by a complex interaction between the selfing rate (s) and the number of alleles (n) at that locus. For $n = 2$ alleles the maximum value of P_a occurs with those two alleles at equal frequencies. For $n = 3$ alleles, the maximum value of P_a also occurs with any two alleles each at a frequency of 0·5 when $s < 1/4$, but with three equally frequent alleles when $s > 1/4$. A final analytical result is that when $s \geqslant 2/3$, the maximum of P_a with $\leqslant n$ alleles occurs with n equally frequent alleles and this maximum monotonically increases with n.

In numerical analyses with $n = 2, 3, \ldots, 100$ alleles and selfing rates $s = 0, 0{\cdot}01, \ldots, 0{\cdot}99, 1$ the maximum of P_a always occurred with $n^* \leqslant n$ equally frequent alleles, but the number of alleles n^* maximizing P_a increases with the selfing rate. For any number of alleles $n \leqslant 100$, the maximum of P_a always occurred with two equally frequent alleles if $s < 1/4$; with three equally frequent alleles if $1/4 < s < 5/11$ (0·25 to 0·45); four equally frequent alleles if $5/11 < s <$ $11/21$ (0·45 to 0·52); five equally frequent alleles if

https://doi.org/10.1017/S0016672301005237 Published online by Cambridge University Press

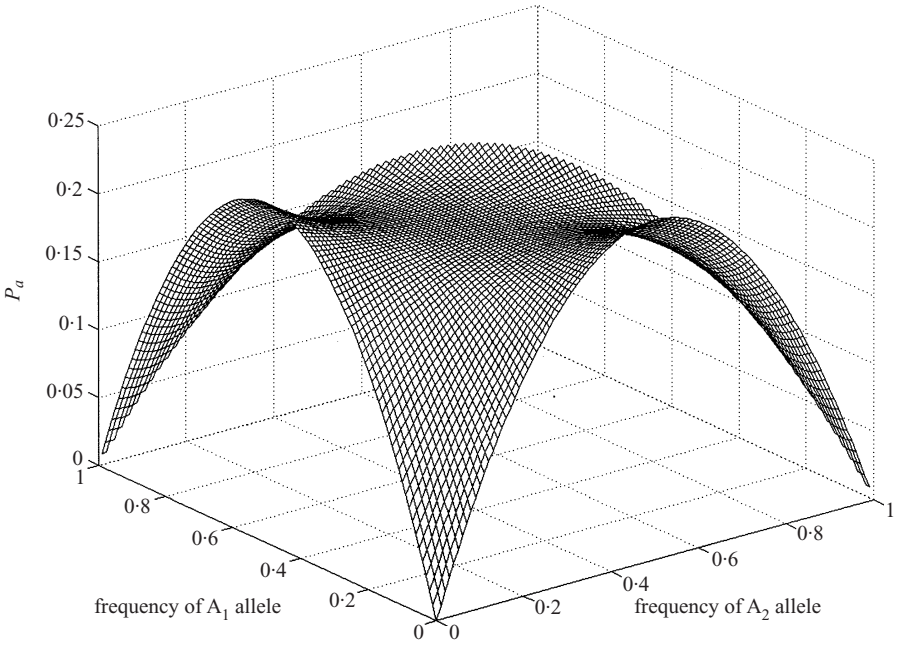

Fig. 2. Probability (P_a from equation (4)) that a randomly chosen offspring has the same heterozygous genotype as the shared parent in a random mating population, for $n = 3$ alleles. The maximum value of 0·25 is approached when any two of the alleles approach equal frequencies of 0·5. The minimum value of 0 is approached when any one of the allele frequencies approaches 1. With three, equally frequent alleles, $P_a = 0{\cdot}22$.

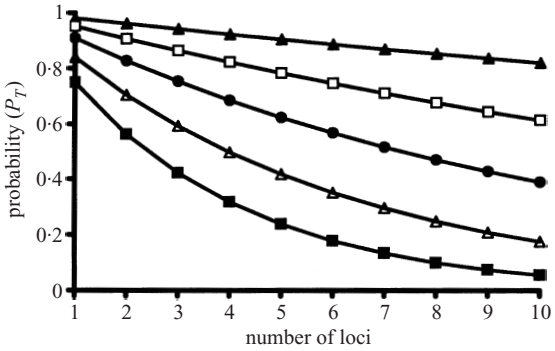

Fig. 3. Probability that a randomly chosen offspring will *not* have the same heterozygous genotype as the shared parent at *any* of the loci in a random mating population, as a function of the number of loci surveyed and the number of equally frequent alleles at each locus. P_T declines as the number of loci surveyed increases or as marker polymorphism decreases.

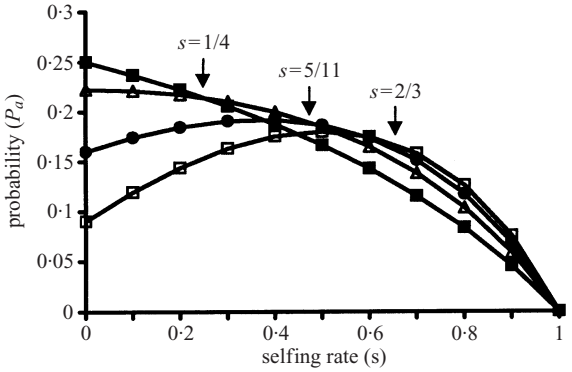

Fig. 4. Probability (P_a from equation (6)) that a randomly chosen offspring will have the same heterozygous genotype as the shared parent in a mixed-mating population as a function of the selfing rate (s). P_a is shown for allele frequency distributions with 2, 3, 5 and 10 equally frequent alleles. With 2 or 3 equally frequent alleles, P_a strictly declines with s. For 5 or 10 equally frequent alleles, P_a first increases but then decreases, as s increases from 0 to 1.

$11/21 < s < 19/34$ (0·52 to 0·56); and so on (see also Fig. 4). Empirically, the maximum of P_a appears to occur with n^* equally frequent alleles, where $s(n^*) < s < s(n^*+1)$ and $s(n)$ is an increasing function of n given by equation (A6) (Appendix B).

(ii) *Effect of selfing rate*

The selfing rate affects both the probability that a given heterozygous female has an offspring of the same heterozygous genotype (P_h) and the probability that a randomly chosen offspring has the same heterozygous genotype as the shared parent (P_a). First consider the effects of selfing rate on P_h (from equation 5). If there are only two alleles at the marker locus, P_h always equals 0·5 regardless of the value of s or the allele frequencies. If there are more than two alleles, P_h, the probability that a heterozygous female has an offspring of the same heterozygous genotype, strictly increases with the selfing rate (s), from $\frac{1}{2}(p_i + p_j)$ when $s = 0$, to 0·5 when $s = 1$. Finally, in the extreme case of a purely selfing population (i.e. $s = 1$), the value of

https://doi.org/10.1017/S0016672301005237 Published online by Cambridge University Press

P_h is independent of the allele frequency distribution and always equals 0·5.

Although P_h strictly increases with increased selfing (for $n \geqslant 3$ alleles), the behaviour of P_a is more complex (see Appendix C). For $n \leqslant 3$ alleles, P_a strictly decreases with increased selfing. For $n \geqslant 4$ alleles, however, the behaviour depends upon the allele frequencies in the population and is specifically determined by the sign of the quantity

$$1-\sum_{i=1}^{n} p_i^2 - 3\sum_{i=1}^{n} p_i^2(1-p_i) = 3\sum_{i<j} p_i p_j(\tfrac{2}{3}-p_i-p_j). \qquad (7)$$

If this quantity is zero or negative, then P_a strictly decreases with s. If, however, this quantity is positive, P_a first increases and then decreases with s, as s increases from 0 to 1 (Fig. 4).

4. Discussion

Extensive efforts, both empirical and theoretical, have focused on estimating the number of parents contributing to half-sib progeny arrays when the genotype of the shared parent is known (Levine *et al*., 1980; Parker & Kornfield, 1996; Kellogg *et al*., 1998; Moran & Garcia-Vazquez, 1998; DeWoody *et al*., 2000; Fiumera *et al*., 2001). Many of the empirical studies count the number of gametotypes contributed by the unshared parents at co-dominant molecular markers by subtracting out the contribution of the shared parent. However, this cannot be done unambiguously when the shared parent and an offspring have the same heterozygous genotype. If this is a common occurrence, estimates of parental numbers may be seriously compromised. To assess how often such ambiguities occur, we have derived and analysed general formulae for co-dominant markers with n alleles for the probability that (1) a *given heterozygous parent* has an offspring with the same heterozygous genotype (P_h) and (2) a *randomly chosen offspring* has the same heterozygous genotype as its shared parent (P_a).

In random mating populations, the maximum of P_a (0·25) occurs with two equally frequent alleles, and P_h is maximized (0·5) when there are only two alleles in the population, whatever their frequencies. The effects of selfing may not be as intuitively obvious. Although P_h is still maximized with two alleles, the allelic distribution which maximizes P_a is governed by a complex interaction between the number of alleles and the selfing rate (s) in the population. Furthermore, while P_h strictly increases with higher selfing rates, P_a can either strictly decrease or first increase with s since it is a convolution of P_h and the equilibrium frequency of heterozygotes (which is a decreasing function of s).

In essence, our probabilities of ambiguity predict the amount of information apt to be gained (or lost) in parentage studies. They should then be expected, in some way or other, to be related to other types of information indices. Therefore, it is not surprising that factors that influence exclusion probabilities (Jamieson & Taylor, 1997) or the PIC index (Botstein *et al*., 1980) also affect the probabilities considered here. For example, in random mating populations, as the number of equally frequent alleles at a given locus increases, both the probability of not being able to exclude a given parent and the probability that a randomly chosen offspring has the same heterozygous genotype as the shared parent (P_a) decrease to 0. In each case, the amount of information expected to be gained increases with increased marker polymorphism.

Given that most parentage studies now employ highly polymorphic molecular markers, P_a is likely to be much lower than the maximum value of 0·25 in random mating populations. It is possible, however, that a large proportion of progeny in any single progeny array might have the same heterozygous genotype as the shared parent. For example, P_h from equation (3) shows that if the shared parent has a heterozygous genotype composed of two relatively common alleles (each at a frequency of 0·20), then on average 20% of her progeny will have her same heterozygous genotype. Such progeny arrays must either be analysed with the reduced information available, which can lead to statistically significant underestimates of parental numbers (Fiumera *et al*., in preparation) or through an alternative method not relying on counting the number of gametotypes. Although explicit likelihood equations may be difficult to formulate (Harshman & Clark, 1998), a likelihood approach based upon offspring genotypes may prove valuable in such cases. An alternative (and potentially preferable) likelihood approach would incorporate P_h and P_a, while still basing estimates of parental numbers on counting the number of distinct gametotypes.

Although initially derived for forensic and parentage inference, these probabilities will probably find application in a wider array of questions in evolutionary genetics. For instance, genetic analysis of progeny arrays may reveal patterns of non-random mating that would not be evident by focusing only on the adult population (e.g. Schoen & Clegg, 1986). Higher proportions of ambiguous offspring than expected under random mating might reflect inbreeding by some parents whereas lower proportions might reflect outbreeding. Without knowing the expected frequency of ambiguous offspring under random mating (i.e. P_a and P_h derived here), it would be impossible to detect deviations with any level of statistical confidence.

In this (and other) practical applications, it is important to recognize that our derivations assume that each offspring was sired by a *random* unshared

https://doi.org/10.1017/S0016672301005237 Published online by Cambridge University Press

parent from the population. In nature, however, most half-sib progeny arrays contain multiple offspring from each unshared parent and therefore each offspring within a given array may not represent an independent sample. This discrepancy can easily be corrected. The observed level of ambiguity can be calculated via Monte Carlo simulations from reduced empirical data sets, each including only a single offspring from each shared parent (repeated thousands of times to generate distributions). The corrected frequency of ambiguous offspring observed in the empirical data would then be compared with the expected frequency (calculated from the probabilities derived here) to test for statistical differences. Lastly, and of considerable practical importance, using P_h and P_a in *a priori* power analyses for parentage and other applications (thus taking into account the information lost from ambiguity) can increase the efficiency of experimental designs, thereby saving limited resources.

Appendix A. Formal derivation of P_h and P_a

(i) P_a *for random mating populations*

Substituting equation (3) and the Hardy–Weinberg genotype frequencies into equation (2) reveals that

$$\begin{aligned} P_a &= \sum_{i<j} P(O=\varphi=A_iA_j) = \sum_{i<j} \tfrac{1}{2}(p_i+p_j)(2p_ip_j) \\ &= \sum_{i<j}(p_i^2p_j+p_ip_j^2) = \sum_{i\neq j} p_i^2p_j = \sum_{i=1}^{n} p_i^2(1-p_i). \end{aligned} \tag{A1}$$

(ii) P_h *and* P_a *for mixed-mating populations*

After substituting in the mixed-mating equilibrium genotype frequencies of males in the population (Marshall & Weir, 1979), equation (1) becomes

$$\begin{aligned} P_h &= \tfrac{1}{2}(1-s)\left[p_i^2+\frac{sp_i(1-p_i)}{2-s}\right] \\ &\quad +\tfrac{1}{2}(1-s)\left[p_j^2+\frac{sp_j(1-p_j)}{2-s}\right]+\tfrac{1}{2}s \\ &\quad +\tfrac{1}{2}(1-s)\left[2p_ip_j-\frac{2sp_ip_j}{2-s}\right] \\ &\quad +\sum_{k\neq i,j}\tfrac{1}{4}(1-s)\left[2p_ip_k-\frac{2sp_ip_k}{2-s}\right] \\ &\quad +\sum_{k\neq i,j}\tfrac{1}{4}(1-s)\left[2p_jp_k-\frac{2sp_jp_k}{2-s}\right] \\ &= \tfrac{1}{2}[s+(1-s)(p_i+p_j)]. \end{aligned} \tag{A2}$$

Substituting equation (A2) and the mixed-mating equilibrium genotype frequencies of heterozygotes (Marshall & Weir, 1979) into equation (2) reduces the formula for P_a to

$$\begin{aligned} P_a &= \sum_{i<j}\tfrac{1}{2}[s+(1-s)(p_i+p_j)]\frac{4(1-s)p_ip_j}{2-s} \\ &= \left(\frac{1-s}{2-s}\right)\left[s\sum_{i<j}2p_ip_j+2(1-s)\sum_{i<j}(p_i+p_j)p_ip_j\right] \\ &= \left(\frac{1-s}{2-s}\right)\left[s\left(1-\sum_{i=1}^{n}p_i^2\right)+2(1-s)\sum_{i=1}^{n}p_i^2(1-p_i)\right]. \end{aligned} \tag{A3}$$

Appendix B. Calculating the maximum value of P_a

Equation (A3) reduces to (A1) when $s=0$ and thus (A3) applies for all $0\leqslant s<1$. Since P_a in equation (A3) is a continuous function of $n-1$ independent variables on a closed bounded region ($p_i\geqslant 0$ for $i=1, 2, \ldots, n-1$; $\sum_{i=1}^{n-1}p_i\leqslant 1$), it has a maximum there. This occurs either on the boundary of the allele frequency space where one or more alleles is absent, or at an admissible critical point inside where all the first partial derivatives are zero.

(i) *Determining the critical points*

For P_a in equation (A3),

$$\frac{\partial P_a}{\partial p_i} = 2\left(\frac{1-s}{2-s}\right)(p_i-p_n)[2-3s-3(1-s)(p_i+p_n)],$$

which equals zero if and only if either $p_i=p_n$ or $p_i=(2-3s)/3(1-s)-p_n$. A potential critical point thus has k of the allele frequencies equalling $p^*=(2-3s)/3(1-s)-p_n$ and the remaining $n-k$ (including p_n) equalling p_n, where $k=0, 1, 2, \ldots, n-1$. Remembering that $p_1+p_2+\cdots+p_n=1$ we have $1=(n-k)p_n+k((2-3s)/3(1-s)-p_n)$, and thus

$$p_n = \frac{3-2k+3s(k-1)}{3(1-s)(n-2k)} \tag{A4}$$

and

$$p^* = \frac{2(n-k)+3s(1-n+k)-3}{3(1-s)(n-2k)}. \tag{A5}$$

The conditions under which the critical points determine valid allele frequencies (all $p_i>0$), as well as the corresponding P_a values, will be treated separately for random mating and mixed-mating populations.

(ii) *Random-mating populations*

If $n=2$, (A4) and (A5) determine a valid critical point with all $p_i>0$ only when $k=0$, for which $p_1=p_2=0{\cdot}5$. If $n\geqslant 3$, then the critical points are valid if and

https://doi.org/10.1017/S0016672301005237 Published online by Cambridge University Press

Table A1. *Critical points and corresponding value of P_a under random mating, given the number of alleles (n) and the valid k value*

n	k values[a]	Critical points	P_a
2	0	$p_1 = p_2 = \frac{1}{2}$	0·25
3	0, 1	$p_1 = p_2 = p_3 = \frac{1}{3}$	0·222
$\geqslant 4$	0	$p_1 = \ldots = p_n = \frac{1}{n}$	$(n-1)/n^2$
	1	$p_n = 1/3(n-2)$, $p^* = (2n-5)/3(n-2)$	$(n-1)(4n-9)/27(n-2)^2$

[a] For $n \geqslant 3$, critical points for $k = 1$ are the same as for $k = n-1$.

only if $k = 0$, 1, or $n-1$ (Table A1); however, since $k = 1$ and $k = n-1$ provide the same critical points, only the cases of $k = 0$ and $k = 1$ are relevant.

We will now demonstrate that for $n \geqslant 2$ alleles, P_a has a maximum value of 0·25, which occurs whenever any two of the alleles have a frequency of 0·5 (i.e. $p_i = p_j = 0{\cdot}5$ for some $i \neq j$ and $p_k = 0$ for $k \neq i, j$). Consider first the cases of $n = 2$ and $n = 3$ alleles. For $n = 2$ alleles, $P_a = 0$ at the two boundary points, $p_1 = 0$ and $p_1 = 1$, and $P_a = 0{\cdot}25$ at the one critical point, $p_1 = p_2 = 0{\cdot}5$. The maximum value of P_a for $n = 2$ alleles is then 0·25 and occurs when $p_1 = p_2 = 0{\cdot}5$. For $n = 3$ alleles, the maximum occurs either on a boundary corresponding to one of the three possible two-allele subsystems (i.e. (i) $p_1 = 0$ and $0 \leqslant p_2 \leqslant 1$; (ii) $p_2 = 0$ and $0 \leqslant p_1 \leqslant 1$; or (iii) $p_2 = 1 - p_1$ and $0 \leqslant p_1 \leqslant 1$) or at the unique internal critical point, $p_1 = p_2 = p_3 = 1/3$. Since $P_a(1/3, 1/3, 1/3) = 0{\cdot}222$, the maximum value of P_a is 0·25 and occurs at the three points on the boundary where $p_i = p_j = 0{\cdot}5$ for some $i \neq j$ and $p_k = 0$ for $k \neq i, j$. Finally, mathematical induction can be used to prove that the maximum value of P_a is 0·25 with two equally frequent alleles for any number of alleles; this follows by showing that for $n \geqslant 4$ alleles the maximum value of P_a at the critical points occurs at those with $k = 1$ and is less than 0·25 for all n.

(iii) *Mixed-mating populations*

The central critical point $p_i \equiv 1/n$ (with $k = 0$) always exists, and equations (A4) and (A5) show that the critical points for $k \geqslant 1$ exist if and only if $s < 2/3$ and either $k < n/2$, $3(1-s)/(2-3s)$, $(2n-3-3s(n-1))/(2-3s)$ or k is greater than each of these three quantities. The number of valid critical points depends upon both the number of alleles (n) and the selfing rate (s). As for random mating, the critical points defined by $k = k^*$ are the same as those for $k = n-k^*$, but with selfing those for $k > 1$ can also be valid (e.g. $k = 2$ for $s = 0{\cdot}4$ and $n = 5$ alleles). For $s \geqslant 2/3$, the only valid critical point is with all alleles at equal frequencies ($p_i \equiv 1/n$).

For $n = 2$ alleles, the maximum occurs at the only valid critical point, where $p_1 = p_2 = 0{\cdot}5$ (with $k = 0$), at which $P_a = \frac{1}{2}((1-s)/(2-s))$ from (6). For $n = 3$ alleles, the critical points defined by $k = 1$ are valid if and only if $s < 1/3$, and at these points $P_a = 2(2-3s)/(9(1-s)(2-s))$, which is always less than $P_a = 2(1-s)(2+s)/9(2-s)$ at the central critical point with $p_1 = p_2 = p_3 = 1/3$ (for $k = 0$). Finally, $P_a(1/2, 1/2) > P_a(1/3, 1/3, 1/3)$ if and only if $s < 1/4$. Therefore, for $n = 3$ alleles, the maximum of P_a occurs at all the boundary points with only two of the alleles present at equal frequency if $s < 1/4$, and with three equally frequent alleles if $s > 1/4$. For $s = 1/4$, P_a is maximized by either two or three equally frequent alleles.

For $n \geqslant 4$ alleles, the analysis is much more complicated. First, when $s \geqslant 2/3$, the only valid critical point is for $k = 0$ when all alleles are equally frequent ($p_i \equiv 1/n$) where $P_a = ((1-s)/(2-s))(n-1)[2+s(n-2)]/(n^2)$. In addition, $P_a(1/n, 1/n, \ldots, 1/n) > P_a(1/(n-1), 1/(n-1), \ldots, 1/(n-1))$ if and only if

$$s > \frac{2(n^2-3n+1)}{(n-2)(3n-1)} = s(n) \tag{A6}$$

where $s(n)$ increases from 1/4 to 2/3 as n increases from 3 to ∞. Thus, if $s \geqslant 2/3$, the maximum of P_a with n alleles occurs at the central critical point ($p_i \equiv 1/n$). Further results for $s < 2/3$ can be obtained numerically as discussed in the text.

Appendix C. The effects of selfing

The effect of s on P_a is determined by the sign of

$$\frac{\partial P_a}{\partial s} = \frac{2A - 3B + 4(B-A)s - (B-A)s^2}{(2-s)^2}$$

where $A = 1 - \sum_{i=1}^{n} p_i^2$, $B = 2\sum_{i=1}^{n} p_i^2(1-p_i)$ and p_i is the frequency of allele A_i. Two different scenarios are possible. If $2A - 3B \leqslant 0$, $\partial P_a/\partial s$ is always negative implying that P_a strictly decreases with s. If $2A - 3B > 0$, $\partial P_a/\partial s$ switches sign from positive to negative implying that P_a initially increases but then decreases with s, as s increases from 0 to 1.

For $n = 2$ alleles, $2A - 3B$ is clearly always $\leqslant 0$ implying that P_a strictly decreases with s, whatever the allele frequencies. We now show that the same is true

https://doi.org/10.1017/S0016672301005237 Published online by Cambridge University Press

for $n = 3$ alleles, by showing that both the minimum and maximum of

$$2A - 3B = 18p_1p_2p_3 - 2(p_1p_2 + p_1p_3 + p_2p_3)$$

are $\leqslant 0$. Since $2A - 3B$ is a continuous function on a closed bounded region ($p_1 \geqslant 0$, $p_2 \geqslant 0$, $p_1 + p_2 \leqslant 1$), the minimum and maximum occur either on the boundary of the admissible allele frequency space or at a valid critical point inside. The latter are the central point $(p_1, p_2) = (1/3, 1,3)$, and the three points $(p_1, p_2) = (1/9, 1/9)$, $(1/9, 7/9)$ or $(7/9, 1/9)$. The maximum value of $2A - 3B$ is thus 0 at $p_1 = p_2 = p_3 = 1/3$ and at the three boundary points where one allele is fixed ($p_1 = 1$ or $p_2 = 1$ or $p_3 = 1$). The minimum value is $-1/2$ and it occurs at the three points on the boundary with two equally frequent alleles [$(p_1, p_2) = (1/2, 1/2)$, $(1/2, 0)$ or $(0, 1/2)$]. This implies that for $n = 3$ alleles, $2A - 3B$ is always $\leqslant 0$ and, therefore, P_a strictly decreases with s for all allele frequency distributions.

For $n \geqslant 4$ alleles the behaviour is more complex because the sign of $2A - 3B$ varies with the allele frequency distribution. For example, with $n \geqslant 4$ equally frequent alleles, $p_i \equiv 1/n < 1/3$ and $2A - 3B = 6\sum_{i<j} p_i p_j (\frac{2}{3} - p_i - p_j) > 0$, and thus P_a first increases and then decreases with s, as s increases from 0 to 1. For other allele frequency distributions, such as four alleles with $p_1 = p_2 = p_3 = \epsilon$ and $p_4 = 1 - 3\epsilon$, $2A - 3B \leqslant 0$ and P_a strictly decreases with s, for $0 \leqslant \epsilon \leqslant 1/8$. The sign-conserving property of continuous functions ensures that this is also true for $n > 4$ alleles near these frequencies. These results imply that for $n \geqslant 4$ alleles both $2A - 3B \leqslant 0$ and $2A - 3B > 0$ occur, along with the two alternative behaviours of P_a, depending upon the actual allele frequencies.

We thank J. Avise, J. A. DeWoody, W. G. Hill, M. Mackiewicz, B. McCoy, D. Pearse, B. Porter, D. Walker and anonymous reviewers for useful comments on the manuscript and A. G. Clark and M. T. Clegg for discussion on estimates of non-random mating. Work was supported by an NIH Training Grant in Genetics (to A.C.F.) and NSF DEB-9906462 (to M.A.A.).

References

Baker, R. J., Makova, K. D. & Chesser, R. K. (1999). Microsatellites indicate a high frequency of multiple paternity in *Apodemus* (Rodentia). *Molecular Ecology* **8**, 107–111.

Bollmer, J. L., Irwin, M. E., Rieder, J. P. & Parker, P. G. (1999). Multiple paternity in loggerhead turtle clutches. *Copeia* **1999**, 475–478.

Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**, 314–331.

Boyd, W. C. (1955). Chances of excluding paternity by the Rh blood groups. *American Journal of Human Genetics* **7**, 229–235.

DeWoody, J. A., DeWoody, Y. D., Fiumera, A. C. & Avise, J. C. (2000). On the number of reproductives contributing to a half-sib progeny array. *Genetical Research* **75**, 95–105.

Evett, I. W. & Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Associates.

Fiumera, A. C., DeWoody, Y. D., DeWoody, J. A., Asmussen, M. A. & Avise, J. C. (2001). Accuracy and precision of methods to estimate the number of parents contributing to a half-sib progeny array. *Journal of Heredity* **92**, 120–126,

Harshman, L. G. & Clark, A. G. (1998). Inference of sperm competition from broods of field-caught *Drosophila*. *Evolution* **52**, 1334–1341.

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

Jones, A. G. & Avise, J. C. (1997). Microsatellite analysis of maternity and the mating system in the gulf pipefish *Syngnathus scovelli*, a species with male pregnancy and sex role reversal. *Molecular Ecology* **6**, 202–213.

Kellogg, K. A., Markert, J. A., Staugger, J. R. & Kocher, T. D. (1998). Intraspecific brood mixing and reduced polyandry in a maternal mouth-brooding cichlid. *Behavioural Ecology* **9**, 309–312.

Kichler, K., Holder, M. T., Davis, S. K., Márquez, M. R. & Owens, D. W. (1999). Detection of multiple paternity in the Kemp's ridley sea turtle with limited sampling. *Molecular Ecology* **8**, 819–830.

Levine, L., Asmussen, M., Olvera, O., Powell, J. R., De La Rosa, M. E., Salceda, V. M., Gaso, M. I., Guzman, J. & Anderson, W. W. (1980). Population genetics of Mexican *Drosophila*. V. A high rate of multiple insemination in a natural population of *Drosophila pseudoobscura*. *American Naturalist* **116**, 493–503.

Marshall, D. R. & Weir, B. S. (1979). Maintenance of genetic variation in apomictic plant populations. *Heredity* **42**, 159–172.

Moran, P. & Garcia-Vazquez, E. (1998). Multiple paternity in Atlantic salmon: a way to maintain genetic variability in relicted populations. *Journal of Heredity* **89**, 551–553.

Parker, A. & Kornfield, I. (1996). Polygynandry in *Pseudotropheus zebra*, a cichlid fish from Lake Malawi. *Environmental Biology of Fishes* **47**, 345–352.

Ross, S. M. (1997). *A First Course in Probability*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

Schoen, D. J. & Clegg, M. T. (1986). Monte-Carlo studies of plant mating system estimation models: the one-pollen parent and mixed mating models. *Genetics* **112**, 927–945.

Weiner, A. S. (1968). Chances of proving non-paternity with a system determined by triple allelic codominant genes. *American Journal of Human Genetics* **20**, 279–282.