Theoretical Population Biology **59**, 315–331 (2001)
doi:10.1006/tpbi.2001.1520, available online at http://www.idealibrary.com on IDEAL®

TPB

A Bayesian Framework for Parentage Analysis: The Value of Genetic and Other Biological Data

Bryan D. Neff

Department of Zoology, University of Western Ontario, London, Ontario N6A 5B7, Canada

Joe Repka

Department of Mathematics, University of Toronto, Toronto, Ontario M5S 3G3, Canada

and

Mart R. Gross

Department of Zoology, University of Toronto, Toronto, Ontario M5S 3G5, Canada

Received November 2, 1999

We develop fractional allocation models and confidence statistics for parentage analysis in mating systems. The models can be used, for example, to estimate the paternities of candidate males when the genetic mother is known or to calculate the parentage of candidate parent pairs when neither is known. The models do not require two implicit assumptions made by previous models, assumptions that are potentially erroneous. First, we provide formulas to calculate the expected parentage, as opposed to using a maximum likelihood algorithm to calculate the most likely parentage. The expected parentage is superior as it does not assume a symmetrical probability distribution of parentage and therefore, unlike the most likely parentage, will be unbiased. Second, we provide a mathematical framework for incorporating additional biological data to estimate the prior probability distribution of parentage. This additional biological data might include behavioral observations during mating or morphological measurements known to correlate with parentage. The value of multiple sources of information is increased accuracy of the estimates. We show that when the prior probability of parentage is known, and the expected parentage is calculated, fractional allocation provides unbiased estimates of the variance in reproductive success, thereby correcting a problem that has previously plagued parentage analyses. We also develop formulas to calculate the confidence interval in the parentage estimates, thus enabling the assessment of precision. These confidence statistics have not previously been available for fractional models. We demonstrate our models with several biological examples based on data from two fish species that we study, coho salmon (*Oncorhychus kisutch*) and bluegill sunfish (*Lepomis macrochirus*). In coho, multiple males compete to fertilize a single female's eggs. We show how behavioral observations taken during spawning can be combined with genetic data to provide an accurate calculation of each male's paternity. In bluegill, multiple males and multiple females may mate in a single nest. For a nest, we calculate the fertilization success and the 95% confidence interval of each candidate parent pair. © 2001 Academic Press

Key Words: **parentage analysis; paternity; maternity; fractional allocation; maximum likelihood; confidence; microsatellite.**

0040-5809/01 $35.00

INTRODUCTION

Parentage analysis in mating systems has traditionally relied on behavioral observations during breeding (Krebs and Davies, 1997). With the advent of molecular techniques in behavioral and evolutionary ecology, genetic markers have gained widespread use and are often used exclusively to assess parentage (for reviews see Queller *et al.*, 1993; Avise, 1994; Jarne and Lagoda, 1996; Petrie and Kempenaers, 1998; Luikart and England, 1999; Sunnucks, 2000). Although both forms of biological data can be informative (e.g., Adams *et al.*, 1992; Philipp and Gross, 1994; Burczyk *et al.*, 1996; Prodöhl *et al.*, 1998; but see Coltman *et al.*, 1999; Smouse *et al.*, 1999), few studies have incorporated both in parentage analyses. In this paper we develop a framework based on genetic likelihood and fractional allocation theory (Devlin *et al.*, 1988; Roeder *et al.*, 1989; Smouse and Meagher, 1994) that allows the incorporation of additional biological information along with genetic marker data in the calculation of parentage and we explore its importance in providing accurate parentage inference. We also provide statistical confidence estimators for the fractional models.

Parentage analysis is required to discern parentage in breeding systems that have multiple mating. Multiple mating occurs when individuals of one sex mate with more than one partner of the opposite sex (Reynolds, 1996). There are two general forms of multiple mating (Neff *et al.*, 2000a): single-sex where there is multiple mating within only one sex, and two-sex where there is multiple mating by both sexes. In single-sex multiple mating, the offspring are produced by a single female that has mated with multiple males or a single male that has mated with multiple females. For example, a female coho salmon (*Oncorhychus kisutch*) may mate with up to four males. In this case, we are interested in determining the proportion of her offspring that are sired by each of the males. In two-sex multiple mating, the offspring are produced by multiple males and multiple females that may each have mated with several individuals. For example, in bluegill sunfish (*Lepomis macrochirus*) the offspring within a nest may be from several females, each of which has mated with the nest-tending parental male as well as other specialized males called cuckolders. In this case, we are interested in determining not only the proportion of offspring sired by each male, but also the proportion sired with each female. Both single-sex and two-sex multiple mating occur widely in nature, and therefore models that allow for each are needed. We have recently developed models and confidence estimators for parentage analysis of offspring produced from single-sex or two-sex multiple mating and when there is incomplete sampling of candidate parents (Neff *et al.*, 2000a, b). Here, we develop models that assume complete sampling of candidate parents.

There are three basic approaches for calculating parentage when all candidate parents have been sampled: (1) exclusion; (2) categorical allocation; and (3) fractional allocation. Exclusion methods (e.g., Ellstrand, 1984) attempt to eliminate all but one of the candidate parents (or parent pairs) based on the available genetic data. When two or more parents cannot be excluded then the assignment of the offspring remains ambiguous. Although potentially an accurate method, exclusion approaches are inefficient, often requiring a large number of loci (e.g., Chakraborty *et al.*, 1988). As such, this approach is generally not feasible, particularly for large studies involving numerous parents and offspring.

Categorical allocation methods (e.g., Meagher, 1986; Meager and Thompson, 1986, 1987) use a log-likelihood ratio, or LOD score, to select the single most likely parent and have the apparent advantage of identifying parent–offspring relationships, which might be useful, for example, when analyzing heritability. Only when two parents have equivalent most likely LOD scores, for example, when they have the same genotypes, does parentage remain ambiguous. A Monte Carlo simulation approach has also been developed to assess the statistical confidence in the categorical assignments (Marshall *et al.*, 1998). However, categorical allocation can systematically overestimate the fertilization success of homozygous individuals while underestimating the success of heterozygous individuals (Devlin *et al.*, 1988; Smouse and Meagher, 1994). This can be particularly problematic when assessing inbreeding depression, since inbred individuals are more homozygous and are therefore favored by the analysis. A reduction in fertilization success resulting from inbreeding depression may subsequently be obscured.

Devlin *et al.* (1988) show that if offspring are instead fractionally allocated among all nonexcluded parents based on their probability of parentage the bias associated with the categorical assignments is circumvented. They provide the first fractional allocation model, which is based solely on transitional (Mendelian) probabilities, and demonstrate how their model can be used to partition individual offspring among candidate parents. Although this initial fractional model is for single-sex multiple mating (i.e., when one genetic parent is known), it is easily amendable to two-sex multiple mating where neither parent is known.

Roeder *et al.* (1989) later develop a maximum likelihood method (similar to fractional models) that incorporates multiple offspring simultaneously in the analysis of parentage (see also Smouse and Meagher, 1994). This approach does not specifically allocate individual offspring, but determines the parentage parameters across all candidate parents that are most likely to have generated the offspring sample (the solution is referred to as the "*mle*"). This approach is superior when the probabilities of parentage of the offspring are not independent, for instance, when the offspring sample is from a single nest where each competing male may have sired multiple offspring. Roeder *et al.* develop models for both single-sex and two-sex multiple mating and provide variance estimators for when the transitional probability matrix is of full column rank (i.e., the genetic contribution of each candidate parent is uniquely identifiable).

Alternatively to modeling the fertilities of individuals, other authors (Burczyk *et al.*, 1996; Smouse *et al.*, 1999) have substituted a function based on biological variables these could include body size of the candidate parents or distances between mates) into fractional or categorical models and use maximum likelihood procedures to solve for the most likely relationships between the variables and the probability of parentage. Instead of calculating reproductive success for each individual, this approach attempts to map the variance in reproductive success directly onto a set of biological variables. This approach may be particularly useful when the goal is to determine the fitness contribution of specific characters, and not individuals.

A potential shortcoming of many parentage models is that they assume a uniform prior probability distribution. That is, *a priori*, they assume that each candidate parent is equally likely to fertilize a given offspring, or that each possible biological function is equally likely. This might be the best assumption in the absence of any information about the actual distribution of the prior probabilities (Devlin *et al.*, 1988). However, it leads to an underestimation of the true variance in reproductive success as individuals with high reproductive success are underestimated while those with low reproductive success are overestimated (Adams *et al.*, 1992; this paper). Examining parentage within a plant population, Adams *et al.* (1992) utilize additional biological data to provide an estimate of the actual prior probability distribution. Using a categorical approach, they show that incorporating the additional data increases the accuracy of their parentage inference.

In this paper we develop a Bayesian framework for parentage analysis that builds on previous models and utilizes genetic and other biological data. Our framework can be used for either single-sex and two-sex multiple mating. Specifically, we provide algorithms to calculate the *expected* fertilities, instead of the *most likely* fertilities calculated by previous models. We show that given the prior probability of parentage, the expected fertilities are unbiased and provide accurate estimates of the true variance in reproductive success. We also show how these algorithms can be used to calculate the confidence in the estimates. These confidence statistics have not previously been available for fractional allocation models. We develop our single-sex and two-sex multiple mating models concurrently since they have analogous structure. The development involves five steps. First, formulas are presented to calculate the transitional probability, a measure of the genetic compatibility of parents and offspring that is based on Mendelian inheritance patterns (see Devlin *et al.*, 1988). Second, we define the parentage vector, which contains the parentage assignments, and the offspring vector, which contains the offspring genetic data (see Roeder *et al.*, 1989). Third, formulas are presented to calculate the probability of a parentage vector based on the transitional probabilities, the offspring vector and, if available, additional biological data (see Adams *et al.*, 1992) such as behavioral observations during breeding. Fourth, we provide methods to determine the most likely and expected parentage vector. Fifth, formulas are presented to determine the confidence in the parentage assignments. Next, we use simulations to examine the effects of key assumptions made by these models. Finally, we present several biological examples to demonstrate the application of the models.

THE MODELS

1. The Transitional Probability

We begin by defining the transitional probabilities for single-sex and two-sex multiple mating. Here we use single-sex to imply that one genetic parent is known and this individual may have mated multiply. Two-sex implies that neither genetic parent is known and each may have mated multiply. Thus, for two-sex multiple mating all combinations of parent pairs must be considered. Note that the single-sex model may be used for mating systems with two-sex multiple mating provided that one genetic parent is known for the entire sample of offspring of interest. In this case, the sample of offspring would have been produced by single-sex multiple mating. For both single-sex and two-sex multiple mating we assume that all candidate parents have been sampled. However, for two-sex multiple mating, we also consider the case where only one sex has been completely sampled.

The transitional probability represents the probability that an individual is an offspring of a putative parent or parent pair and is based on Mendelian inheritance. It equals zero if, at any locus, either the mother or the father can be excluded as a potential genetic parent. When loci are linked, the transitional probability can be calculated with modification (see Devlin *et al.*, 1988). Linkage reduces the effective number of loci and the precision of parentage analysis, but itself can be used to infer additional dimensions of relatedness such as differentiating grandmother–granddaughter from aunt–niece (Thompson and Meagher, 1998). Sample calculations of transitional probabilities for single-sex and two-sex multiple mating are provided in Tables I and II, respectively. Similar tables are presented in other papers (e.g., Meagher, 1986; Marshall *et al.*, 1998).

(a) *Complete sampling of candidate parents from both sexes* (*single-sex or two-sex multiple mating*). Suppose that we have a set of M putative mothers, F putative fathers, and C unique offspring genotypes. There is a total of $A = M \cdot F$ parent pairs and $A \cdot C$ mother–father–offspring combinations. Note that for single-sex multiple mating either M or F equals 1. For each of these triplets the transitional probability can be calculated from

$$T_{ac} = T(g_c \mid g_m, g_f) = \prod_{l=1}^{L} \left(\sum_{i=1}^{2} \sum_{j=1}^{2} t_{lij} \right), \qquad (1)$$

where T_{ac} is the transitional probability for the a parent pair (consisting of the m mother and f father) and c offspring triplet; t_{lij} equals 0.25 if the combination of the m mother's allele i and the f father's allele j is equivalent to the c offspring's genotype at locus l; otherwise t_{lij} equals 0; g is the multilocus genotype of the m mother, f father, or c offspring; and L is the total number of loci utilized to genotype the triplet.

TABLE I

Sample Calculations of the Transitional Probability (T_{ac}) for Single-Sex Multiple Mating.

Genotypes			
Genetic parent	Putative parent	Offspring	T_{ac}
BB	*BB*	*BB*	1
BB	*BC*	*BC*	1/2
BC	*BB*	*BC*	1/2
BC	*BC*	*BC*	1/2
BB	*BB*	*BC*	0

Note. The genetic parent refers to the known parent that has multiple mates. Genotypes are represented by uppercase letters.

TABLE II

Sample Calculations of the Transitional Probability (T_{ac}) for Two-Sex Multiple Mating Given Complete Sampling of Both Sexes or Only One Sex.

Genotypes				
Parent(s)	Offspring	S_l	F_l	T_{ac}
Both sexes sampled				
$BB \times BB$	*BB*	—	—	1
$BB \times BC$	*BB*	—	—	1/2
$BC \times BC$	*BC*	—	—	1/2
$BC \times BC$	*BB*	—	—	1/4
$BB \times BB$	*CC*	—	—	0
Only one sex sampled				
BB	*BC*	1	c	c
BC	*BB*	1/2	b	$1/2 \cdot b$
BC	*BC*	1	$1/2 \cdot (b+c)$	$1/2 \cdot (b+c)$
BB	*BB*	1	b	b
BB	*CC*	0	—	0

Note. Genotypes are represented by uppercase letters and the population allele frequencies are represented by corresponding lowercase letters. All other variables are defined in the text.

(b) *Complete sampling of only one sex* (*two-sex multiple mating only*). Suppose that we have sampled only the M putative mothers or F putative fathers. There is a total of $A = M$ or $A = F$ putative parents and $A \cdot C$ parent–offspring combinations. For each of these parent–offspring pairs the transitional probability can be calculated from

$$T_{ac} = T(g_c \mid g_a) = \prod_{l=1}^{L} (S_l \cdot F_l), \qquad (2)$$

where T_{ac} is the transitional probability for parent a and offspring c; F_l equals the frequency of the c offspring's allele at locus l when it is homozygous and shares the allele with parent a; F_l equals the average frequency of the c offspring's alleles when it is heterozygous and shares both alleles with parent a; or F_l equals the frequency of the c offspring's unshared allele when it shares exactly one allele with parent a; g is the multilocus genotype of the a parent and c offspring; L is the total number of loci utilized to genotype the parents and offspring; and S_l is the proportion of the a parent's alleles that are shared by offspring c.

The transitional probabilities associated with all combinations of candidate parents and offspring are then grouped into the **T** matrix, which has rows that correspond to each parent or parent pair (indexed by a and hereafter referred to as parent) and columns that correspond to each unique offspring genotype (indexed by c). **T** has A rows and C columns and is defined by

$$\mathbf{T} = \begin{vmatrix} T_{11} & \cdots & T_{1C} \\ \vdots & \ddots & \vdots \\ T_{A1} & \cdots & T_{AC} \end{vmatrix}. \tag{3}$$

The element of row a and column c (T_{ac}) is therefore the transitional probability associated with the a parent and the c unique offspring genotype and is calculated using either (1) or (2). Devlin *et al.* (1988) used the transitional probabilities exclusively to partition individual offspring among candidate parents.

2. The Parentage and Offspring Vectors

Next, following Roeder *et al.* (1989), define the parentage vector $\mathbf{Par} = (\text{Par}_1, \text{Par}_2, ..., \text{Par}_A)$ with elements representing the potential parentage assignments, expressed as a proportion of the total offspring sample, of each of the candidate parents. **Par** therefore has A elements, which correspond to the rows in **T**. Also define the offspring vector $\mathbf{X} = (X_1, X_2, ..., X_C)$ with elements representing the number of offspring in the sample that have genotype g_c. **X** therefore has C elements, which correspond to the columns in **T**.

Given **T** and **X**, the parentage of each parent or parent pair could be calculated using standard matrix algebra by solving the equation $\mathbf{Par} \cdot \mathbf{T} = \mathbf{X}$ (Schoen and Stewart, 1986; Roeder *et al.*, 1989). However, since biological systems have random variation in Mendelian inheritance (following the multinomial distribution) this is generally impossible (i.e., it is unlikely that **X** will fall in the row space of **T**). Instead, all possible parentage assignments to the A parents (i.e., all possible parentage vectors) must be considered. From Bayes' rule we can calculate the probability of a particular parentage vector **Par** based on the transitional probability matrix **T**, the offspring vector **X**, and, if available, additional biological data. We formalize the calculations below.

3. The Probability of a Parentage Vector

The probability of **Par** given **X** can be calculated from

$$\Pr(\mathbf{Par} \mid \mathbf{X}) = \frac{\Pr(\mathbf{X} \mid \mathbf{Par}) \cdot \Pr(\mathbf{Par})}{\Pr(\mathbf{X})}. \tag{4}$$

The probability of **X**, Pr(**X**), can be calculated from the multinomial theorem. However, it is unnecessary to determine it here since it is independent of **Par** and will later become part of a normalization constant. Based on the multinomial theorem the probability of **X** given **Par** (sometimes called the likelihood) can be calculated from (modified from Roeder *et al.*, 1989)

$$\Pr(\mathbf{X} \mid \mathbf{Par}) = \left(\frac{N!}{\prod_i (X_i!)} \right) \cdot \prod_{c=1}^{C} \Psi_c^{X_c}, \tag{5}$$

where

$$\mathbf{\Psi} = \mathbf{Par} \circ \mathbf{T} \tag{6}$$

i.e.,

$$\Psi_c = \text{Par}_1 \cdot T_{1c} + \text{Par}_2 \cdot T_{2c} + \cdots + \text{Par}_A \cdot T_{Ac}.$$

The final component of (4) is Pr(**Par**) which represents the prior probability of a given **Par** vector (i.e., independent of the genetic data). In previous fractional allocation models it has been assumed that this probability follows a uniform distribution (see Devlin *et al.*, 1988; Pena and Chakraborty, 1994; Smouse and Meagher, 1994), and therefore, *a priori* all parentage vectors are equally likely. Hence Pr(**Par**) is a constant and like Pr(**X**) becomes part of the normalization constant. In this case, Pr(**Par** | **X**) is calculated directly from Pr(**X** | **Par**). However, the actual distribution of Pr(**Par**) depends, in part, on the dynamics of the mating system, such as whether all males are equally likely to be genetic fathers. As an example, suppose that in a mating system males compete to control a harem of females and only the dominant male will reproduce with them. In this case, Pr(**Par**) does not follow a uniform distribution. Instead, the probability of all parentage vectors that do not assign the parentage of all offspring from the harem to a single male should be zero. Assuming that Pr(**Par**) follows a uniform distribution will lead to biased parentage inference and underestimates the variance in reproductive success. We quantify the effects of this assumption below under Assumption Violation.

Adams *et al.* (1992) show that additional biological data can be used to estimate the actual prior probability distribution. Although they provide a function specific to their mating system, their analysis can be easily generalized. Suppose that the biological data are in the form of a set of traits **t** that relate to parentage. The distribution $\Pr_{\mathbf{B}}(\mathbf{Par})$ (we use the subscript "B" to denote the additional biological data used to estimate the prior probability) can be expressed as

$$\Pr_{\mathbf{B}}(\mathbf{Par}) = F(\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \ldots), \qquad (7)$$

where $\mathbf{t}_i$ is a vector containing A elements reflecting the states of a biological trait i for each of the A parents, and F is a function relating these traits to the prior probability of **Par**. Here we assume that each trait $\mathbf{t}_i$ has a finite number of states (i.e., values that the trait can take on) and that the traits are independent. Accordingly, $\Pr_{\mathbf{B}}(\mathbf{Par})$ can be calculated from

$$\Pr_{\mathbf{B}}(\mathbf{Par}) = k \cdot \prod_{ia} \Pr(\mathrm{Par}_a \mid t_{ia}), \qquad (8)$$

where k is the normalization constant and t_{ia} represents the value of trait i for parent a. The product is calculated over all traits and parents.

Categorizing each trait into a finite set of states does not exclude traits that have a continuous distribution. For example, if for trait i there is a linear relationship between $\mathbf{t}_i$ and $\Pr_{\mathbf{B}}(\mathbf{Par})$, then the confidence interval around the linear regression can be used to calculate $\Pr(\mathrm{Par}_a \mid t_{ia})$ in (8). If multiple dependent biological traits are used then statistical procedures such as principal components analysis can generate a set of independent traits. Essentially, we are after a function that accurately relates a set of biological traits to the probability of parentage (for specific examples see Adams *et al.*, 1992; Burczyk *et al.*, 1996; Smouse *et al.*, 1999).

4. The Most Likely and Expected Parentage Vector

Given the formula for the probability of a particular parentage vector (Eq. 4), the most likely and expected parentage vectors can be calculated. The most likely parentage vector, which is calculated by all previous models, maximizes $\Pr(\mathbf{Par} \mid \mathbf{X})$ [or $\Pr(\mathbf{X} \mid \mathbf{Par})$] and represents the single most probable parentage assignments. In Appendix 1 we present an iterative algorithm that solves for the maximum likelihood solution. However, the most likely parentage vector assumes that $\Pr(\mathbf{Par} \mid \mathbf{X})$ is symmetrically distributed about this value and will otherwise provide biased parentage estimates. The expected parentage vector represents an average of all possible vectors weighted by their associated probabilities and provides unbiased estimates of parentage independent of the symmetry of the distribution of $\Pr(\mathbf{Par} \mid \mathbf{X})$. The expected parentage vector can be calculated from

$$\overline{\mathbf{Par}} = k \cdot \int_{\mathbf{Par}} (\Pr(\mathbf{Par} \mid \mathbf{X}) \cdot \mathbf{Par})\, d\mathbf{Par}; \qquad (9)$$

here k is the normalization constant defined such that $\int \Pr(\mathbf{Par} \mid \mathbf{X}) = 1$.

Equation (9) can be computationally difficult to calculate, particularly when there are a large number of candidate parents that are not excluded. We have therefore developed methods that convert the integrand to a simpler algebraic equation. From (4), we see that the integrand involves the polynomial (5), a constant $\Pr(\mathbf{X})$, and another function, either $\Pr(\mathbf{Par})$ or $\Pr_{\mathbf{B}}(\mathbf{Par})$. If $\Pr(\mathbf{Par})$ or $\Pr_{\mathbf{B}}(\mathbf{Par})$ is a polynomial, then the integrand is a polynomial in the variables $\mathrm{Par}_1, \ldots, \mathrm{Par}_A$ and can be calculated using the results of Theorem 1 in Appendix 2, which shows how to evaluate the integral for any monomial. If there are large numbers of offspring and nonexcluded parents, the polynomials will involve large numbers of monomials and the computation may be time-consuming. However, this approach is far more tractable than the original integral formula in (9). Furthermore, Monte Carlo simulation approaches (e.g., Manly, 1997) may be able to accurately estimate (9) when it is not possible to evaluate directly. By the Stone–Weierstrass theorem, the functions $\Pr(\mathbf{Par})$ and $\Pr_{\mathbf{B}}(\mathbf{Par})$ can always be approximated by polynomials to within any desired precision, so this requirement does not limit the applicability of the technique.

In the case of two-sex multiple mating and complete sampling of both sexes, the expected paternity or maternity of individual parents can be calculated by adding all elements of the expected parentage vector that include the desired parent.

5. Confidence Intervals

Confidence intervals can be established for the estimates based on (4). First, generate the probability distribution of Par_a given $\mathbf{X}$ for $\mathrm{Par}_a \in \{0, 1\}$ from

$$\Pr(\mathrm{Par}_a = \lambda \mid \mathbf{X}) = k \cdot \int_{\substack{\mathbf{Par} \\ \mathrm{Par}_a = \lambda}} \Pr(\mathbf{Par} \mid \mathbf{X})\, d\mathbf{Par}, \qquad (10)$$

where k is the normalization constant. Next, determine the values of Par_a (denoted below with an asterisk) that cut off upper and lower "tails" of areas $1-\alpha/2$ and $\alpha/2$, respectively, by solving

$$k \cdot \int_{\mathrm{Par}_a=0}^{\mathrm{Par}_a^*} \int_{\mathbf{Par}} \Pr(\mathbf{Par} \mid \mathbf{X})\, d\mathrm{Par}_a\, d\mathbf{Par} = \frac{\alpha}{2} \text{ or } 1-\frac{\alpha}{2}. \quad (11)$$

As an example, for the 95% confidence interval, α equals 0.05, and (11) is solved for values of Par_a that cut off the lower and upper 2.5% of the normalized $\Pr(\mathbf{Par} \mid \mathbf{X})$ distribution. In the case of two-sex multiple mating and complete sampling of both sexes, the confidence interval can also be calculated for their individual paternity or maternity estimates. First, let **J** be the set containing each element of the **Par** vector that includes the putative parent. The confidence interval for the parent's cumulative success (i.e., with all of its mates in **J**) is found by solving

$$k \cdot \int_{\substack{\mathrm{Par}_a \\ a \in \mathbf{J}}}^{\sum \mathrm{Par}_a \leqslant \mathrm{Par}^*} \int_{\substack{\mathrm{Par}_b \\ b \notin \mathbf{J}}}^{\sum \mathrm{Par}_b = 1-\sum \mathrm{Par}_a} \times \Pr(\mathbf{Par} \mid \mathbf{X})\, d\mathbf{Par} = \frac{\alpha}{2} \text{ or } 1-\frac{\alpha}{2}. \quad (12)$$

Again, when a large number of parents must be considered the integrand in (11) or (12) can be expressed as a polynomial for which Theorem 2 in Appendix 2 can be used to solve each monomial. In the case of (12), the monomials associated with the inner integral must first be solved, followed by the outer integral.

ASSUMPTION VIOLATIONS

(a) *The Prior Probability Distribution* Pr(Par)

To test the effects of assuming that Pr(**Par**) follows a uniform distribution, we considered a simple case where there are two males competing for a female's eggs (the general results are applicable to any number of males and females). We assumed that the prior paternity of Male 1 could follow one of seven probability distributions (Table III). As an example, the uniform distribution implies that Male 1 is just as likely to have a parentage of 0 or 0.35 or 1 or any other value between 0 and 1. This could represent a mating system where parentage is random. In contrast, the normal distribution implies that Male 1 is most likely to have a parentage of 0.5 and is very unlikely to have a parentage of 0 or 1. This distribution could represent a mating system where there is little variance in male quality and it is likely that each male gets close to the same paternity. The seven distributions cover several general types of mating dynamics that exist in nature and thus should provide a reasonable description of the various outcomes possible for distributions of Pr(**Par**). Nearly all previous models have assumed the uniform distribution in their analyses (for an exception see Adams *et al.*, 1992). First, we considered each of the seven distributions of Pr(**Par**) and examined their individual effects of assuming that it follows one of the other six distributions. We calculated the bias (accuracy) and variance (precision) in the parentage estimates [as calculated from (9)] resulting from the assumption. Ideally, both should be minimized. The bias was calculated from

$$\text{bias} = \int_{\mathrm{Par}_1=0}^{1} (|\Pr_{\mathrm{act}}(\mathrm{Par}_1) - \Pr_{\mathrm{est}}(\mathrm{Par}_1)|) \times \Pr_{\mathrm{act}}(\mathrm{Par}_1)\, d\mathrm{Par}_1, \quad (13)$$

where $\Pr_{\mathrm{act}}(\mathrm{Par}_1)$ and $\Pr_{\mathrm{est}}(\mathrm{Par}_1)$ represent the actual and estimated distribution of $\Pr(\mathrm{Par}_1)$, respectively. The variance was calculated from

$$\text{variance} = \int_{\mathrm{Par}_1=0}^{1} (\Pr_{\mathrm{act}}(\mathrm{Par}_1) - \Pr_{\mathrm{est}}(\mathrm{Par}_1))^2 \times \Pr_{\mathrm{act}}(\mathrm{Par}_1)\, d\mathrm{Par}_1. \quad (14)$$

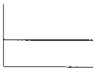
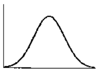
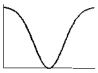
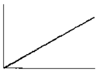
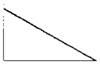
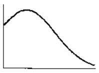
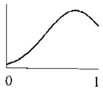
Table IV summarizes the average bias and variance introduced for each combination of distributions. When the correct distribution of Pr(**Par**) was used, the bias and variance were both zero. Overall, the uniform distribution performed the best, minimizing both the bias and the variance, and the normal distribution performed the least well. Although in each case the uniform distribution was not the best alternative to the actual distribution, unlike the other distributions, it performed nearly as well as the best alternative for every distribution and was therefore the most robust. Generally, in the absence of any information on the distribution of Pr(**Par**), the uniform distribution is the best *a priori* assumption. However, if the actual distribution of Pr(**Par**) is not uniform, then this assumption will generate biased parentage inference (also see below). If additional biological data are available that can reliably estimate the distribution of Pr(**Par**), it should be used to provide more accurate parentage inferences.

(b) *Variance in Reproductive Success*

To test the effects of calculating the most likely versus expected parentage vector and assuming a uniform prior

TABLE III

Summary of the Functions Used to Examine the Effects of Estimating Pr(Par) without Biological Data

Function			
Graph	Name	Equation	Biological description
	Uniform	$f(x) = 1$	Any parentage equally likely
	Normal	$f(x) = 2.4 \times e^{(-18 \times (x-0.5)^2)}$	Equal parentage most likely; all or no parentage highly unlikely
	Inverted normal	$f(x) = 1.715 \times (1 - e^{(-18 \times (x-0.5)^2)})$	All or no parentage most likely; equal parentage highly unlikely
	Increasing linear	$f(x) = 2x$	All parentage most likely; no parentage highly unlikely
	Decreasing linear	$f(x) = 2 - 2x$	No parentage most likely; all parentage highly unlikely
	Skewed right	$f(x) = 1.625 \times e^{(-5 \times (x-0.25)^2)}$	Low parentage most likely; all parentage highly unlikely
	Skewed left	$f(x) = 1.625 \times e^{(-5 \times (x-0.75)^2)}$	High parentage most likely; no parentage highly unlikely

Note. The graphs plot Pr(**Par**) versus Par_1 (over 0, 1). Each function is graphed and its name, equation, and biological description are given.

TABLE IV

Summary of the Effects of Assuming that Pr(Par) Follows one of Seven Distributions

	Uniform	Normal	Inverted normal	Increasing	Decreasing	Skewed right	Skewed left
Uniform	0.0(0.0)	0.75(0.70)	0.48(0.27)	0.50(0.33)	0.50(0.33)	0.48(0.28)	0.48(0.28)
Normal	0.86(0.93)	0.0(0.0)	1.2(1.6)	0.91(1.0)	0.91(1.0)	0.81(0.81)	0.81(0.81)
Inverted normal	0.54(0.36)	1.5(2.7)	0.0(0.0)	0.50(0.37)	0.50(0.37)	0.65(0.60)	0.65(0.60)
Increasing	0.50(0.33)	1.1(1.5)	0.70(0.77)	0.0(0.0)	1.7(2.3)	1.0(1.3)	0.23(0.09)
Decreasing	0.50(0.33)	1.1(1.5)	0.70(0.77)	1.7(2.3)	0.0(0.0)	0.23(0.09)	1.0(1.3)
Skewed right	0.44(0.23)	0.87(0.93)	0.73(0.81)	0.89(0.95)	0.21(0.07)	0.0(0.0)	0.90(0.95)
Skewed left	0.44(0.23)	0.87(0.93)	0.73(0.81)	0.21(0.07)	0.89(0.95)	0.90(0.95)	0.0(0.0)
Total	3.3(2.4)	6.2(8.3)	4.5(5.0)	4.7(5.0)	4.7(5.0)	4.1(4.0)	4.1(4.0)

Note. The net bias is presented with the variance in parentheses. The total indicates the sum of the values within the column. Overall, the uniform distribution had the least bias and variance. The actual distribution $\text{Pr}_{\text{act}}(\textbf{Par})$ is presented down the left side and the estimated distribution $\text{Pr}_{\text{est}}(\textbf{Par})$ across the top. See Table III for a description of the distributions.

probability distribution on the variance in reproductive success we performed the following simulation. We modeled pairs of individuals, a dominant and a subordinate, that compete for a set of 10 or 20 offspring. To introduce a skew in reproductive success we assumed that the paternity of a dominant followed the prior probability distribution of $\Pr(\mathrm{Par}_1) = 2 \times \mathrm{Par}_1$. Therefore, on average a dominant fertilizes two-thirds of the offspring, but in any given pair could fertilize more or less. The subordinate in each pair fertilizes the remaining offspring. For a given pair, genotypes were generated for each male and one female based on a single locus with five equally common alleles. The paternity of the dominant was then randomly drawn from the prior probability distribution. Based on the paternity value, each of the offspring was probabilistically assigned to either the dominant or the subordinate. When an offspring was assigned to a parent its genotype was generated from the genetic father and mother based on Mendelian inheritance. Equations (4) and (9), with either a uniform or the correct skewed prior probability distribution, were then used to calculate the most likely and expected paternity for the dominant and subordinate. Each estimate was compared to the actual paternities and the difference (bias) was recorded. The procedure was repeated for 10,000 pairs.

On average the dominant male fertilized two-thirds and the subordinate male one-third of the offspring, as expected based on the prior probability distribution. However, when a uniform prior probability distribution was assumed, the paternity estimates were biased (Table V). The paternity of the dominant was underestimated, while the paternity of the subordinate was overestimated. This resulted in an underestimation of the true reproductive skew. This bias was lower for the expected paternity compared to the most likely paternity and decreased when a larger number of offspring were analyzed. When the correct prior probability distribution was incorporated into the analysis, the expected paternities were unbiased. Interestingly, the most likely paternity now overestimated the success of the dominant and therefore overestimated reproductive skew. This bias is attributed to the skew in the parentage distribution $\Pr(\mathbf{Par} \mid \mathbf{X})$ (also see the biological examples below). It cannot be attributed to row dependence in the transitional probability matrix such as when the two males had equivalent genotypes, since in these cases correct paternities would be assigned based solely on the prior probabilities. The bias decreased with increasing numbers of alleles or loci, but increased with increasing numbers of candidate parents (data not shown). Overall, the expected parentage vector provides less biased parentage estimates compared to the most likely parentage vector and provides unbiased estimates when the true prior probability distribution is known. In this latter case, the expected parentage vector (Eq. 9) provides unbiased estimates of the variance in reproductive success and reproductive skew.

EXAMPLES

To demonstrate the methods we consider three biological examples: (1) single-sex multiple mating; (2) two-sex multiple mating with complete sampling of both sexes; and (3) two-sex multiple mating with complete sampling of only one sex. While these examples are based on two fish species that we study, coho salmon (*O. kisutch*) and bluegill sunfish (*L. macrochirus*), they are meant to broadly demonstrate the application of the models.

Single-Sex Multiple Mating

In this example males compete for dominance and form a mating hierarchy, as in the case of many salmonids (e.g.,

TABLE V

Effects of a Uniform Prior Probability Distribution and Most Likely versus Expected Vectors on the Accuracy of Parentage Inference

Number offspring	Paternity (%)		Uniform prior probability[a]		Actual prior probability[a, b]	
	Dominant	Subordinate	Most likely	Expected	Most likely	Expected
10	66.7	33.3	61.4 (5.0)	61.7 (1.8)	73.9 (2.2)	66.7 (1.9)
20	66.7	33.3	61.7 (4.3)	63.0 (1.2)	71.7 (1.7)	66.7 (1.1)

[a] Values represent mean paternity (%) for the dominant male with variance in parentheses.

[b] Actual prior probability distribution was $\Pr(\mathrm{Par}_1) = 2 \times \mathrm{Par}_1$ (see text).

Fleming and Gross, 1994; Quinn, 1999). Imagine that two males were observed mating with a single female and Male 1 obtained the dominant (alpha) mating position in the hierarchy and Male 2 obtained the subordinate (beta) position. Ten offspring were later collected from the female's nest and we wanted to calculate the paternity and a confidence interval for each male. Genotypes were therefore obtained at a single locus for each male, the female, and the 10 offspring (e.g., Neff *et al.*, 2000c; Table VI). The dominant male had a positional advantage during mating and therefore was expected to fertilize a greater proportion of the eggs (Gross, Neff, and Fleming, in review). Suppose that the probability distribution of paternity given the hierarchy position of Male 1 was expected to follow the following normalized distribution:

$$\mathrm{Pr_B}(\mathrm{Par}_1) = 2 \cdot \mathrm{Par}. \qquad (15)$$

First, we calculated the most likely and expected paternity of the two males based on (4) and (9) and using only the genetic data, the transitional probabilities for single-sex multiple mating (e.g., Table I), and assuming that Pr(**Par**) followed the uniform distribution. We also calculated the 95% confidence interval associated with the paternity estimates using (11).

In this example the parentage vector consists of two elements $\mathbf{Par} = (\mathrm{Par}_1, \mathrm{Par}_2 = 1 - \mathrm{Par}_1)$ representing the parentage of Male 1 and Male 2 with the female, respectively. The offspring vector has two elements **X** = (6, 4) representing the six *Aa* and four *aa* offspring (see Table VI), and the **T** matrix is

$$\mathbf{T} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{vmatrix}. \qquad (16)$$

The distribution of Pr(**Par** | **X**), based only on the genetic data, over the possible values of **Par**, is presented in Fig. 1. Pr(**Par** | **X**) is maximized at **Par** = (0.8, 0.2) (Fig. 1a). However, since the distribution is not

TABLE VI

Summary of the Genetic Data for the Biological Examples

Individual	**G**	**X**$_i$
Single-sex multiple mating		
Female	*aa*	—
Male 1	*Aa*	—
Male 2	*AA*	—
Offspring	*Aa*	6
	aa	4
Two-sex multiple mating		
Female 1	*AC*	—
Female 2	*AB*	—
Male 1	*AA*	—
Male 2	*BB*	—
Offspring	*AA*	8
	AB	4
	AC	6
	BC	2

Note. The genotypes (**G**) are given for the offspring and the candidate parents. X_i represents the number of offspring with the corresponding genotype.

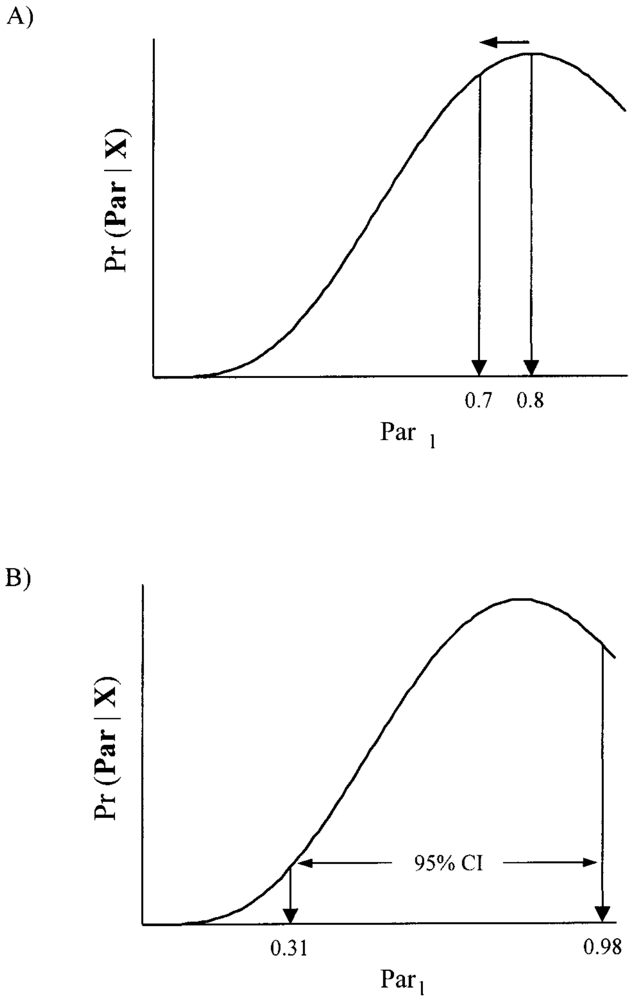


FIG. 1. Calculation of the paternity and confidence intervals for Male 1 in the first biological example. Here it assumed that Pr(**Par**) follows the uniform distribution. (a) The expected paternity is lower than the most likely paternity since the distribution of Pr(**Par** | **X**) is skewed. (b) The 95% confidence interval in the paternity estimate is calculated by determining the values of Par$_1$ that represent the lower and upper 2.5% of the area of the Pr(**Par** | **X**) distribution. In this example, the most likely estimate for the paternity of Male 1 is 80% while the expected paternity is 70% with a 95% confidence interval of 31–98%.

symmetrically distributed about the most likely value, it provides a biased estimate of each male's parentage. Since the distribution is skewed toward lower parentage, the expected parentage for the first male is less than the most likely value, and consequently, the parentage for the second male is more than the most likely value. In this example, the expected parentage for the two males is $\mathbf{Par} = (0.7, 0.3)$. That is, the expected parentage for Male 1 is 70 % and for Male 2 it is 30 %.

The 95 % confidence interval in the parentage estimate was also calculated from the distribution of $\Pr(\mathbf{Par} \mid \mathbf{X})$ and is 0.31–0.98 for Male 1 and 0.02–0.69 for Male 2 (Fig. 1B). The confidence interval is not evenly distributed about the expected or most likely value, again since the $\Pr(\mathbf{Par} \mid \mathbf{X})$ distribution is asymmetrical about these values. Since we had only two males in this example, there is only a single independent parentage (e.g., the parentage of Male 2 is simply what remains from Male 1: $\mathrm{Par}_2 = 1 - \mathrm{Par}_1$). Therefore, the most likely or expected parentages for the two males sum to 1 and similarly, the upper bound of the confidence interval for one male and the lower bound of the other male sum to 1. We arbitrarily chose to display the distribution for Male 1. Either increasing the number of loci used to genotype the parents and offspring or increasing the number of offspring sampled from the female's nest would decrease the confidence interval and therefore increase our certainty in the parentage estimates (data not shown).

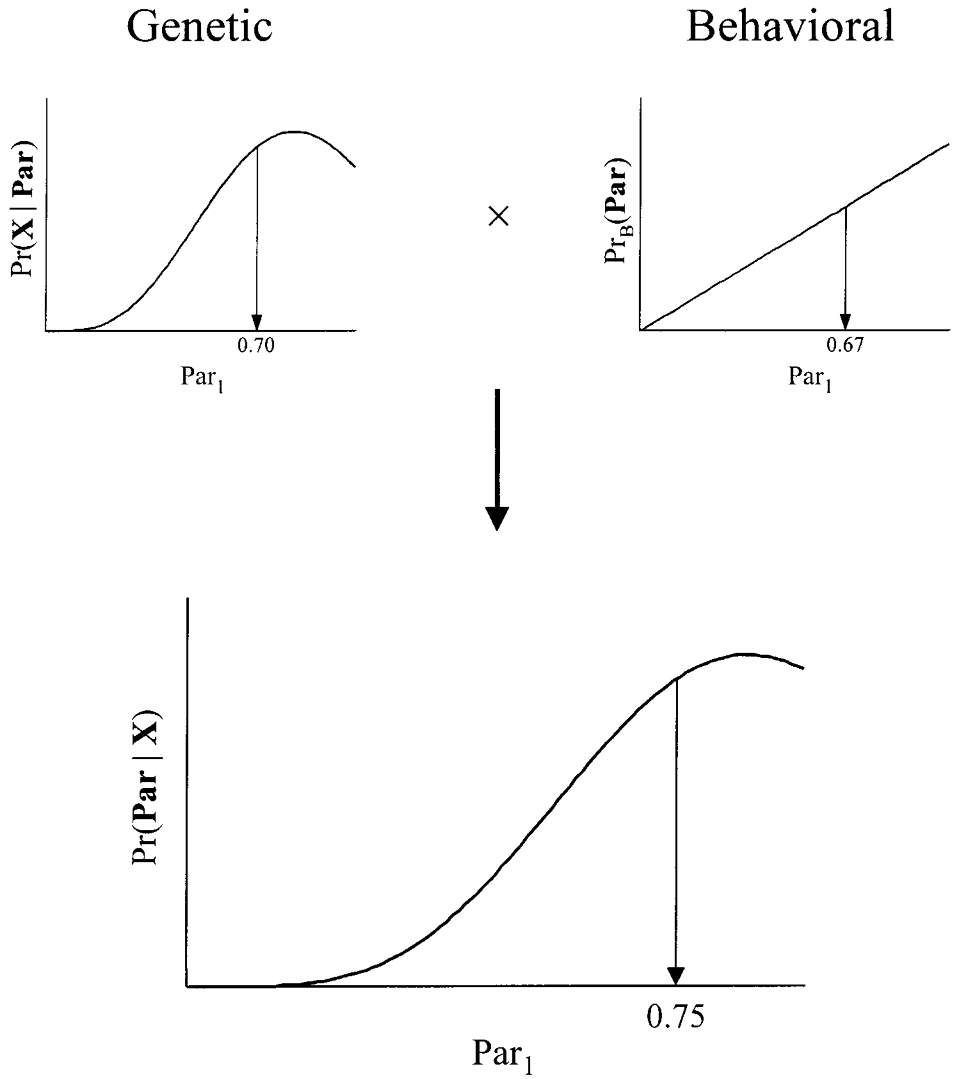


FIG. 2. Calculation of the paternity and confidence intervals for Male 1 in the first biological example. Here both genetic ($\Pr(\mathbf{X} \mid \mathbf{Par})$) and behavioral data ($\Pr_B(\mathbf{Par})$) are included in the analysis of parentage, providing a more accurate estimate of paternity. In this example the expected paternity for Male 1 is 75 %.

Next, we repeated the analysis including the biological data (i.e., the behavioral observation of hierarchy position; Eq. 15). The distribution of $\Pr(\mathbf{Par} \mid \mathbf{X})$, based on both the genetic and the biological data, over the possible values of **Par** is presented in Fig. 2. In this case, the expected parentage is $\mathbf{Par} = (0.75, 0.25)$. The 95% confidence interval for Male 1 is 0.38–0.99 and for Male 2 is 0.01–0.62. Since Male 1 obtained a superior position in the mating hierarchy his expected paternity increased compared to the analysis including only the genetic data. Similarly, since Male 2 obtained an inferior position, his paternity decreased. Therefore, based only on the genetic data, the paternity estimates would have been biased. Furthermore, assuming a uniform prior probability distribution would have underestimated the true variance in reproductive success.

Two-Sex Multiple Mating

In this example imagine that males and females both mate with multiple partners, and the eggs are laid in a communal nest, such as in many fish (e.g., Gross, 1982) and birds (e.g., Macedo and Bianchi, 1997). Genotypes were obtained at a locus for 20 offspring from the nest and for 2 males and 2 females found in the vicinity of the nest (Table IV). We wanted to calculate the parentage and a confidence interval for each of the parent pairs. Since we had no knowledge of the likely distribution of $\Pr(\mathbf{Par})$ we assumed that it follows a uniform distribution.

First, we calculated the expected parentage of each of the four parent pairs, as well as their individual paternity or maternity, based on (4) and (9) and the transitional probabilities for two-sex multiple mating and complete sampling of both sexes (e.g., Table II). We also calculated the 95% confidence interval associated with the parentage estimates, as well as the confidence interval associated with the individual paternity and maternity estimates.

In this case the parentage vector consists of four elements $\mathbf{Par} = (\mathrm{Par}_1, \mathrm{Par}_2, \mathrm{Par}_3, \mathrm{Par}_4 = 1 - \mathrm{Par}_1 - \mathrm{Par}_2 - \mathrm{Par}_3)$ representing the parentage of Male 1 with Female 1, Male 1 with Female 2, Male 2 with Female 1, and Male 2 with Female 2, respectively. The offspring vector has four elements $\mathbf{X} = (8, 4, 6, 2)$ representing the *AA*, *AB*, *AC*, and *BC* offspring, respectively (see Table VI), and the **T** matrix is

$$\mathbf{T} = \begin{vmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 \end{vmatrix}. \tag{17}$$

The expected parentage (95% CI) of Male 1 and Female 1 is 0.54 (0.29–0.73), that of Male 1 and Female 2 is 0.19 (0.01–0.48), that of Male 2 and Female 1 is 0.20 (0.05–0.41), and that of Male 2 and Female 2 is 0.07 (0.00–0.22). Individually, the expected paternity of Male 1 is 0.73 (0.52–0.94) and that of Male 2 is 0.27 (0.14–0.37). The expected maternity of Female 1 is 0.74 (0.56–0.91) and that of Female 2 is 0.26 (0.10–0.40).

Next, we repeated the analysis assuming that only the females are collected and calculated the maternity and a confidence interval for each. In this case, the **Par** vector would consist of only two elements $\mathbf{Par} = (\mathrm{Par}_1, \mathrm{Par}_2 = 1 - \mathrm{Par}_1)$ representing the maternity of Female 1 and Female 2, respectively. The offspring vector has the same four elements $\mathbf{X} = (8, 4, 6, 2)$, but here the **T** matrix is

$$\mathbf{T} = \begin{vmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \end{vmatrix}. \tag{18}$$

The expected maternity (95% CI) of Female 1 is 0.60 (0.07–0.98) and that of Female 2 is 0.40 (0.02–0.93). Compared to the previous analysis that included the genotypes of the putative fathers, the maternity estimates have a much larger confidence interval and therefore are considerably less precise.

In these examples for simplicity we have considered only a single locus with relatively low resolving power and therefore the confidence intervals are abnormally large. Generally, we have found that with only a few loci better estimates are obtained with considerably narrower confidence intervals (data not shown).

DISCUSSION

In this paper we have developed fractional allocation models for mating systems with single-sex or two-sex multiple mating. These models build on fractional allocation and genetic likelihood theory (Devlin *et al.*, 1988; Roeder *et al.*, 1989; Smouse and Meagher, 1994) and allow for both genetic and other biological data to be utilized in the parentage inference. We have also developed confidence estimators for our models, thus enabling the analysis of the precision of the estimates.

Previous fractional allocation models make two implicit assumptions that can lead to inaccurate parentage inferences. First, they assume that parentage is random and therefore the prior probability of parentage is uniformly distributed. Categorical allocation models (and many other likelihood models such as individual (R) and population (F_{st}) relatedness estimators) make a similar assumption. We have shown that, in the absence of additional information to the genetic data, assuming that the prior probability follows the uniform distribution is the best *a priori* assumption as it minimizes both the bias and the variance in the estimates of fertilization success (also see Devlin *et al.*, 1998). However, this assumption can lead to biased estimates and particularly underestimates the variance in reproductive success. In turn, this can result in underestimates of, for example, reproductive skew, selection coefficients, and the effective size of populations. We have therefore developed a model based on the framework of Adams *et al.* (1992) that incorporates other biological data, for example, behavioral observations during mating, to estimate the prior probability distribution and thereby provide more accurate parentage inference.

Second, previous fractional allocation models assume that the probability of parentage based on the genetic data [$\Pr(\mathbf{Par} \mid \mathbf{X})$] is symmetrically distributed about the most likely parentage. However, for most cases in nature this distribution will be skewed (e.g., see the biological examples above), and therefore the most likely value provides a biased estimate of the expected parentage. In contrast to the first assumption, the most likely parentage will generally overestimate major contributors and underestimate minor contributors. This can lead to an overestimate of the variance in reproductive success. The model that we have developed to calculate the expected parentage, in conjunction with the correct prior probability distribution, can provide unbiased parentage inference.

Assessing the statistical confidence in parentage estimates is an important component of parentage analyses (e.g., Pena and Chakraborty, 1994; Evett and Weir, 1998; Marshall *et al.*, 1998). While confidence estimators in parentage assignments for categorical allocation models are available through a simulation method developed by Marshall *et al.* (1998), no such methods have previously been developed for fractional allocation models. Our statistics calculate the confidence interval for an individual's parentage expressed as a proportion. For example, an individual's expected parentage may be 75% with a 95% confidence interval of 70–80% These statistics now make it possible to calculate statistical confidence in parentage estimates based on fractional allocation models.

Roeder *et al.* (1989) show that modeling parentage based on a set of offspring simultaneously can provide more accurate inference compared to examining individual offspring. This approach assumes that the parentage of one offspring is related to the parentage of other offspring in the sample. Thus, it is useful to identify the correct level of analysis with respect to the offspring sample. For example, consider the mating system of the bluegill sunfish. Parental male bluegill make nests in a colony and spawn with multiple females, during which time specialized cuckolder males sneak fertilizations (Gross, 1982). Females may spawn eggs in multiple nests and cuckolders may intrude in multiple nests. Thus, while a parental male's fertility is limited to his own nest, a cuckolder's and female's may be distributed in multiple nests throughout the colony. As such, the best level of analysis may be to model the entire colony (i.e., to consider the offspring from every nest in the colony simultaneously) as opposed to individual nests or individual offspring. At the colony level, the prior probability distribution may be less skewed and easier to estimate, thus providing more accurate and precise parentage inference. The specific effects of the level of analysis on parentage inference are likely to be complex and are beyond the scope of this paper, but could provide additional insights worth investigating.

It has been proposed that categorical allocation has the advantage of identifying parent–offspring links, which can be useful when calculating, for example, narrow sense heritability (e.g., Marshall *et al.*, 1998). The fractional allocation models developed here can also identify the most likely parent for individual offspring. In this case, the analysis is performed using only a single focal offspring (also see Devlin *et al.*, 1988) and the most likely parent is identified as having the highest expected parentage. The expected parentage also reflects the confidence in the assignment and should be similar to the categorical confidence statistic developed by Marshall *et al.* (1998). Our model has the added advantage of including alternatives to the uniform distribution for the prior probability of parentage, but does not consider incomplete sampling of candidate parents or genotyping errors as does the model of Marshall *et al.* The confidence interval provided by our model can also be used to determine whether the most likely parent is significantly more likely than another parent, for example, if the 95% confidence intervals do not overlap. This interval reflects the certainty in the confidence estimate and is similar to power in other statistics (e.g., see Zar, 1999).

To demonstrate the application of our fractional allocation model to the assignment of an individual offspring to a single parent, consider the following example. Suppose that we assign an offspring among three putative fathers and the expected parentage vector is $\mathbf{Par} = (0.80, 0.20, 0)$. Since only a single offspring is in the analysis, the parentage values represent the probability of paternity for each of the three males. Therefore, in this example the first male is the most likely father of the offspring. Further, he is four times more likely to be the father than the second male. The third male must have been excluded by the genetic data ($\text{Par}_3 = 0$) and therefore it is not possible that he is the father of the offspring. Suppose that the 95% confidence intervals associated with the first two males' paternities are 0.70–0.90 and 0.15–0.25, respectively. Since the distributions do not overlap, we can conclude that the first male is significantly more likely than the second male to be the genetic father of the offspring. Although it may be desirable to identify parent–offspring relationships in this manner, if applied repeatedly to several offspring, it will lead to inaccurate parentage inferences similarly to categorical allocation models (Devlin *et al.*, 1988).

In many cases it is unnecessary to identify parent–offspring relationships. As an example, suppose that we were interested in determining the heritability of a trait based on several offspring and a set of candidate parents that we had collected. Conventionally, once the parent–offspring relationships are established a regression is performed on the trait values of the offspring and parents (e.g., Falconer, 1989). However, given that fractional allocation probabilistically assigns offspring among putative parents and does not identify a single parent, an index must be calculated for the parent trait value. This value can be calculated simply from the average of the trait values from each nonexcluded parent weighted by their probability of parentage. For example, given two nonexcluded putative parents with probabilities of parentage of 0.80 and 0.20, the trait value would be calculated as 80% of the first male's plus 20% of the second male's. This indexed trait value could then be used in the conventional regression and should provide a more accurate calculation of heritability.

Complete sampling of both sexes in mating systems with two-sex multiple mating provides considerably more precise estimates of parentage. We presented biological examples for two sampling scenarios of two-sex multiple mating: complete sampling of both sexes and complete sampling of only one sex. The fractional allocation model provides equally accurate estimates of parentage under both of these scenarios. That is, whether both sexes or only one sex have been completely sampled, the estimates are equally unbiased. However, including genetic data from both sexes increases the precision of the estimates. Researchers may wish to consider the tradeoff between sampling requirements, such as complete sampling of both sexes, or increasing the number of loci, when determining the optimal approach to a desired level of confidence.

Traditionally, genetic data have been used to confirm or refute mating system dynamics inferred from behavioral observations (e.g., Gibbs *et al.*, 1990; Philipp and Gross, 1994; Petrie and Kempenaers, 1998). However, both genetic and other biological data can be informative and often complement one another (e.g., Adams *et al.*, 1992; Philipp and Gross, 1994; but see Coltman *et al.*, 1999). Our model, which incorporates both biological and genetic data into the fractional allocation of offspring, utilizes a greater amount of potentially available data and therefore increases the accuracy and efficiency of parentage analyses. Finally, we have also developed formulas to calculate the genetic likelihood $[\Pr(\mathbf{X} \mid \mathbf{Par})]$ when only some of the candidate parents have been sampled (Neff *et al.*, 2000b). These latter formulas can be incorporated into the current Bayesian models. Thus, the models presented here could provide a general framework for parentage analysis.

APPENDIX 1

Here we present an iterative algorithm that solves for the most likely parentage vector, based on the approach outlined in Smouse and Meagher (1994).

First, set each element of the parentage vector **Par** equal: $\mathbf{Par} = (A^{-1}, ..., A^{-1})$. Next, use the following equation to calculate new values for each element of **Par**,

$$\text{Par}_a^{i+1} = \sum_{c=1}^{C} \left(\frac{\text{Par}_a^i \cdot \Pr(\text{Par}_a^i) \cdot T_{ac}}{\Delta_c} \cdot \frac{X_c}{N} \right),$$

where

$$\Delta_c = \Sigma_{a=1}^{A} \left(\text{Par}_a^i \cdot \Pr(\text{Par}_a^i) \cdot T_{ac} \right).$$

Iterate by incrementing i by 1 and recalculating each element of **Par**. Repeat this process until the values of **Par** converge (i.e., when $\mathbf{Par}^{i+1} = \mathbf{Par}^i$). Provided that the **T** matrix is of maximal rank, the converged values of **Par** (denoted below by an asterisk) will provide a global maximum for $\Pr(\mathbf{Par} \mid \mathbf{X})$ (see Smouse and Meagher, 1994 and references within). From these values the paternity and maternity for each father and mother in the respective sets F and M can be calculated from

$$\text{Pat}_f = \sum_{a \in f} \text{Par}_a^*;$$

$$\text{Mat}_m = \sum_{a \in m} \text{Par}_a^*.$$

APPENDIX 2

Here we derive two theorems that convert a monomial integrand to an algebraic equation. The first theorem is used to solve (9) and the second theorem to solve (11) or (12). Here, for clarity, we use p_i to represent the elements of **Par** (i.e., $p_i = \mathrm{Par}_i$).

Let Δ_n be the $(n-1)$-dimensional set

$$\Delta_n = \left\{ (p_1, p_2, ..., p_n) \colon p_i \geqslant 0, \text{ for all } i, \sum_{i=1}^{n} p_i = 1 \right\}.$$

THEOREM 1. *If $x_1, ..., x_n$ are nonnegative integers, then*

$$\iint \cdots \int_{\Delta_n} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n} \, dp_1 \, dp_2 \cdots dp_{n-1} = \frac{x_1! \, x_2! \cdots x_n!}{(x_1 + x_2 + \cdots + x_n + n - 1)!}.$$

Proof. First, when $n = 2$,

$$\int_{\Delta_2} p^x q^y \, dp = \int_0^1 p^x (1-p)^y \, dp.$$

Using integration by parts, with $u = p^x$, $v' = (1-p)^y$, we find this equals

$$-\left(\frac{1}{y+1} p^x (1-p)^{y+1} \right) \Bigg|_0^1 + \frac{x}{y+1} \int_0^1 p^{x-1} (1-p)^{y+1} \, dp = \frac{x}{y+1} \int_0^1 p^{x-1} (1-p)^{y+1} \, dp.$$

By induction, we find that

$$\int_0^1 p^x (1-p)^y \, dp = \frac{x(x-1)(x-2) \cdots 1}{(y+1)(y+2) \cdots (x+y)} \int_0^1 (1-p)^{x+y} \, dp = \frac{x! \, y!}{(x+y+1)!},$$

as required.

The general case is proved by induction on n; assuming the result is known for lower dimensions, write

$$\begin{aligned}
&\iint \cdots \int_{\Delta_n} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n} \, dp_1 \, dp_2 \cdots dp_{n-1} \\
&= \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{n-2}} p_1^{x_1} p_2^{x_2} \cdots p_{n-1}^{x_{n-1}} \\
&\quad \times (1 - p_1 - \cdots - p_{n-1})^{x_n} \, dp_1 \, dp_2 \cdots dp_{n-1} \\
&= \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{n-2}} p_1^{x_1} p_2^{x_2} \cdots p_{n-1}^{x_{n-1}} \\
&\quad \times (1 - p_1 - \cdots - p_{n-2})^{x_n} \\
&\quad \times \left(1 - \frac{p_{n-1}}{1 - p_1 - \cdots - p_{n-2}} \right)^{x_n} \\
&\quad \times dp_1 \, dp_2 \cdots dp_{n-1}.
\end{aligned}$$

Making a change of variables, taking p_{n-1} to $\frac{p_{n-1}}{1 - p_1 - \cdots - p_{n-2}}$, we get

$$\begin{aligned}
&\int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{n-3}} \int_0^1 p_1^{x_1} p_2^{x_2} \cdots p_{n-1}^{x_{n-1}} \\
&\quad \times (1 - p_1 - \cdots - p_{n-2})^{x_n + x_{n-1} + 1} \\
&\quad \times (1 - p_{n-1})^{x_n} \, dp_1 \, dp_2 \cdots dp_{n-1} \\
&= \iint \cdots \int_{\Delta_{n-1}} p_1^{x_1} p_2^{x_2} \cdots p_{n-2}^{x_{n-2}} \\
&\quad \times (1 - p_1 - \cdots - p_{n-2})^{x_{n-1} + x_n + 1} \\
&\quad \times dp_1 \, dp_2 \cdots dp_{n-2} \\
&\quad \times \int_0^1 p_{n-1}^{x_{n-1}} (1 - p_{n-1})^{x_n} \, dp_{n-1}.
\end{aligned}$$

The last integral can be evaluated using the result from above for $n = 2$ and the previous one by the induction hypothesis. The result is

$$\begin{aligned}
&\frac{x_1! \, x_2! \cdots x_{n-2}! \, (x_n + x_{n-1} + 1)!}{(x_1 + x_2 + \cdots + x_{n-1} + x_n + 1 + (n-1) - 1)!} \\
&\quad \times \frac{x_{n-1}! \, x_n!}{(x_{n-1} + x_n + 1)!} \\
&= \frac{x_1! \, x_2! \cdots x_{n-2}! \, (x_n + x_{n-1} + 1)!}{(x_1 + x_2 + \cdots + x_{n-1} + x_n + n - 1)!},
\end{aligned}$$

as required.

It will also be necessary to evaluate the integral with one coordinate held constant. If $1 \leqslant k \leqslant n$, $0 \leqslant c \leqslant 1$, let

$$\Delta_{n-1}^{k}(c) = \left\{ (p_1, p_2, ..., p_n) \colon p_k = c,\ p_i \geqslant 0, \text{ for all } i, \right.$$
$$\left. \text{and } \sum_{i=1}^{n} p_i = 1 \right\}.$$

THEOREM 2. *If $x_1, ..., x_n$ are nonnegative integers, then*

$$\iint \cdots \int_{\Delta_{n-1}^{k}(c)} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}\, dp_1\, dp_2 \cdots \times dp_{k-1}\, dp_{k+1} \cdots dp_{n-1}$$
$$= c^{x_k} (1-c)^{x_1 + x_2 + \cdots + x_{k-1} + x_{k+1} + \cdots + x_n + n - 2} \times \frac{x_1!\, x_2! \cdots x_{k-1}!\, x_{k+1}! \cdots x_n!}{(x_1 + x_2 + \cdots + x_{k-1} + x_{k+1} + \cdots + x_n + n - 2)!}.$$

Proof. First note that

$$\iint \cdots \int_{\Delta_{n-1}^{k}(c)} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}\, dp_1\, dp_2 \cdots \times dp_{k-1}\, dp_{k+1} \cdots dp_{n-1}$$
$$= c^{x_k} \iint \cdots \int_{\Delta_{n-1}^{k}(1-c)} p_1^{x_1} p_2^{x_2} \cdots p_{k-1}^{x_{k-1}} p_{k+1}^{x_{k+1}} \cdots \times p_n^{x_n}\, dp_1\, dp_2 \cdots dp_{k-1}\, dp_{k+1} \cdots dp_{n-1}.$$

A change of variables $p_i \mapsto (1-c)\, q_i$ transforms the integral over $(p_1, p_2, ..., p_{k-1}, c, p_{k+1}, ..., p_n) \in \Delta_{n-1}^{k}(1-c)$ into an integral over $(q_1, q_2, ..., q_{k-1}, q_{k+1}, ..., q_n) \in \Delta_{n-1}$. The integral becomes

$$c^{x_k} \iint \cdots \int_{\Delta_{n-1}} ((1-c)\, q_1)^{x_1} ((1-c)\, q_2)^{x_2} \cdots \times ((1-c)\, q_{k-1})^{x_{k-1}} ((1-c)\, q_{k+1})^{x_{k+1}} \cdots \times ((1-c)\, q_n)^{x_n} (1-c)^{n-2}\, dq_1\, dq_2 \cdots dq_{k-1} \times dq_{k+1} \cdots dq_{n-1}$$
$$= c^{x_k} (1-c)^{x_1 + x_2 + \cdots + x_{k-1} + x_{k+1} + \cdots + x_n + n - 2} \times \iint \cdots \int_{\Delta_{n-1}} q_1^{x_1} q_2^{x_2} \cdots q_{k-1}^{x_{k-1}} q_{k+1}^{x_{k+1}} \cdots q_n^{x_n} \times dq_1\, dq_2 \cdots dq_{k-1}\, dq_{k+1} \cdots dq_{n-1}.$$

This last integral can be evaluated using Theorem 1 with $n-1$ in place of n, giving

$$c^{x_k} (1-c)^{x_1 + x_2 + \cdots + x_{k-1} + x_{k+1} + \cdots + x_n + n - 2} \times \frac{x_1!\, x_2! \cdots x_{k-1}!\, x_{k+1}! \cdots x_n!}{(x_1 + x_2 + \cdots + x_{k-1} + x_{k+1} + \cdots + x_n + n - 2)!},$$

as required.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their helpful comments. This work was supported by the NSERC of Canada.

REFERENCES

Adams, W. T., Griffin, A. R., and Moran, G. F. 1992. Using paternity analysis to measure effective pollen dispersal in plant populations, *Am. Nat.* **140**, 762–780.

Avise, J. C. 1994. “Molecular Markers, Natural History and Evolution,” Chapman & Hall, New York.

Burczyk, J., Adams, W. T., and Shimizu, J. Y. 1996. Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus attenuata* Lemmon.) stand, *Heredity* **77**, 251–260.

Chakraborty, R., Meagher, T. R., and Smouse, P. E. 1988. Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity, *Genetics* **118**, 527–536.

Coltman, D. W., Bancroft, D. R., Robertson, A., Smith, J. A., Clutton-Brock, T. H., and Pemberton, J. M. 1999. Male reproductive success in a promiscuous mammal: Behavioural estimates compared with genetic paternity, *Mol. Ecol.* **8**, 1199–1209.

Devlin, D., Roeder, K., and Ellstrand, N. C. 1988. Fractional paternity assignment: Theoretical development and comparison to other methods, *Theor. Appl. Genet.* **76**, 369–380.

Ellstrand, N. C. 1984. Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. *Am. Nat.* **126**, 606–612.

Evett, I. W., and Weir, B. S. 1998. “Interpreting DNA Evidence,” Sinauer, Sunderland, MA.

Falconer, D. S. 1989. “An Introduction to Quantitative Genetics,” 3rd ed., Longman and Wiley, New York.

Fleming, I. A., and Gross, M. R. 1994. Breeding competition in a Pacific salmon (Coho: *Oncorhynchus kisutch*): Measures of natural and sexual selection, *Evolution* **48**, 637–657.

Gibbs, H. L., Weatherhead, P. J., Boag, P. T., White, B. N., Tabak, L. M. and Hoysak, D. J. 1990. Realized reproductive success of polygynous red-winged blackbirds revealed by DNA markers, *Science* **250**, 1394–1397.

Gross, M. R. 1982. Sneakers, satellites and parentals: Polymorphic mating strategies in North American sunfishes, *Z. Tierpsychol.* **60**, 1–26.

Jarne, P., and Lagoda, J. L. P. 1996. Microsatellites, from molecules to populations and back, *Trends Ecol. Evol.* **8**, 285–288.

Krebs, J. R., and Davies, N. B. (Eds.) 1997. "Behavioural Ecology: An Evolutionary Approach," 4th ed., Blackwell Sci., Oxford.

Luikart, G., and England, P. R. 1999. Statistical analysis of microsatellite DNA data, *Trends Ecol. Evol.* **14**, 253–256.

Macedo, R. H., and Bianchi, C. A. 1997. Communal breeding in tropical Guira Cuckoos Guira guira: Sociality in the absence of a saturated habitat, *J. Avian Biol.* **28**, 207–215.

Manly, B. F. J. 1997. "Randomization, Bootstrapping and Monte Carlo Methods in Biology," Chapman & Hall, New York.

Marshall, T. C., Slate, J., Kruuk, L. E. B., and Pemberton, J. M. 1998. Statistical confidence for likelihood-based paternity inference in natural populations, *Mol. Ecol.* **7**, 639–655.

Meagher, T. R. 1986. Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most likely male parents, *Am. Nat*. **128**, 199–215.

Meagher, T. R., and Thompson, E. 1986. The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction, *Theor. Popul. Biol*. **29**, 87–106.

Meagher, T. R., and Thompson, E. 1987. Analysis of parentage for naturally established seedlings of *Chamaelirium luteum* (Liliaceae), *Ecology* **68**, 803–812.

Neff, B. D., Repka, J., and Gross, M. R. 2000a. Parentage analysis with incomplete sampling of candidate parents and offspring, *Mol. Ecol.* **9**, 515–528.

Neff, B. D., Repka, J., and Gross, M. R. 2000b. Statistical confidence in parentage analysis with incomplete sampling: How many loci and offspring are needed?, *Mol. Ecol.* **9**, 529–539.

Neff, B. D., Fu, P., and Gross, M. R. 2000c. Microsatellite multiplexing in fish, *Trans. Am. Fish. Soc.* **129**, 584–593.

Pena, S. D. J., and Chakraborty, R. 1994. Paternity testing in the DNA era, *Trends Genet*. **10**, 204–209.

Petrie, M., and Kempenaers, B. 1998. Extra-pair paternity in birds: Explaining variation between species and populations, *Trends Ecol. Evol*. **13**, 52–58.

Philipp, D. P., and Gross, M. R. 1994. Genetic evidence of cuckoldry in bluegill, *Lepomis macrochirus*. *Mol. Ecol.* **3**, 563–569.

Prodöhl, P. A., Loughry, W. J., McDonough, C. M., Nelson, W. S., Thompson, E. A., and Avise, J. C. 1998. Genetic maternity and paternity in a local population of Armadillos assessed by microsatellite DNA markers and field data, *Am. Nat*. **151**, 7–19.

Queller, D. C., Strassmann, J. E., and Hughes, C. R. 1993. Microsatellites and kinship, *Trends Ecol. Evol*. **8**, 285–288.

Quinn, T. P. 1999. Variation in Pacific salmon reproductive behaviour associated with species, sex and levels of competition, *Behaviour* **136**, 179–204.

Reynolds, J. D. 1996. Animal breeding systems, *Trends Ecol. Evol*. **11**, 68–72.

Roeder, K., Devlin, B., and Lindsay, B. G. 1989. Applications of maximum likelihood methods to population genetic data for the estimation of individual fertilities, *Biometrics* **45**, 363–379.

Schoen, D. J., and Stewart, S. C. 1986. Variation in male reproductive investment and male reproductive success in white spruce, *Evolution* **40**, 1109–1120.

Smouse, P. E., and Meagher, T. R. 1994. Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (*Liliaceae*), *Genetics* **136**, 313–322.

Smouse, P. E., Meagher, T. R., and Kobak, C. J. 1999. Parentage analysis in *Chamaelirium luteum* (L.) Gray (Liliaceae): Why do some males have higher reproductive contributions?, *J. Evol. Biol.* **12**, 1069–1077.

Sunnucks, P. 2000. Efficient genetic markers for population biology, *Trends Ecol. Evol.* **15**, 199–203.

Thompson, E. A., and Meagher, T. R. 1998. Genetic linkage in the estimation of pairwise relationship, *Theor. Appl. Genet.* **97**, 857–864.

Zar, J. H. 1999. "Biostatistical Analysis," 4th ed., Prentice–Hall, Simon & Schuster, Upper Saddle River, NJ.