Parentage analysis with incomplete sampling of candidate parents and offspring

B. D. NEFF, * J. REPKA† and M. R. GROSS*

*Department of Zoology, †Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 3G5

Abstract

Many breeding systems include 'multiple mating' in which males or females mate with multiple partners. We identify two forms of multiple mating: 'single-sex', where the next-generation individuals (NGIs) are the product of multiple mating by one sex; and 'two-sex', where the NGIs are the product of multiple mating by both sexes. For both mating systems we develop models that estimate the proportion of NGIs that is fathered (paternity) or mothered (maternity) by the putative parents. The models only require genetic data from the parent or parents in question and the sample of NGIs, as well as an estimate of population allele frequencies. The models provide unbiased estimates, can accommodate loci with many alleles and are robust to violations of their assumptions. They allow researchers to address intractable problems such as the parentage of seeds found on the ground, juvenile fish in a stream, and nestlings in a communal breeding bird. We demonstrate the models using genetic data from a nest of the bluegill sunfish *Lepomis macrochirus*, where the NGIs may be from multiple females that have spawned with multiple males from different life histories (cuckolder and parental).

Introduction

Genetic markers are revolutionizing the study of wild populations. An important area for applying genetic markers is the calculation of parentage, the genetic relationship between candidate parents and the next-generation individuals (NGIs). Parentage allows the potential assessment of reproductive success, social structure, kinship and, ultimately, fitness (see Avise 1994; Jarne & Lagoda 1996; Petrie & Kempenaers 1998; for reviews). However, there is a need for mathematical models that apply genetic markers to parentage calculations.

The aim of this work is to provide models for calculating parentage, and its components of maternity and paternity, in breeding systems with multiple mating. Multiple mating occurs when individuals of one sex mate with more than one partner of the opposite sex (Reynolds 1996). We identify two general forms of multiple mating: single-sex and two-sex. In single-sex multiple mating, the NGIs 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 fish may spawn with multiple males, or she may be in a harem of multiple females that spawn with a single male. In two-sex multiple mating, the NGIs are produced by multiple males and multiple females that have mated with several members of the opposite sex. For example, juvenile fish in a stream may come from many females and males that have each spawned with several members of the opposite sex. Both forms of multiple mating occur widely in nature (Reynolds 1996) and therefore models that allow for each are needed.

Current studies of parentage in breeding systems with multiple mating have focused on assigning NGIs among a set of candidate parents using exclusion, categorical and fractional allocation models (e.g. Ellstrand 1984; Meagher 1986; Meagher & Thompson 1986, 1987; Chakraborty *et al.* 1988; Devlin *et al.* 1988; Morin *et al.* 1994; Smouse & Meagher 1994; Keane *et al.* 1997; Coltman *et al.* 1998; Prodöhl *et al.* 1998). These models allow calculations of parentage when all candidate parents have been sampled. By contrast, Westneat *et al.* (1987) and Wrege & Emlen (1987) have developed models that do not require the sampling of all candidate parents, but their models provide

Correspondence: Bryan D. Neff. Fax: 416-978-8532; E-mail: neff@zoo.utoronto.ca

Table 1 Overview of the models. All models require information on the allele frequencies of the breeding population, and a sample of the next-generation individuals (NGIs)

Model name	Mating system	Question	Additional information needed
Two-Sex Multiple Mating			
Two-Sex Paternity	Multiple genetic mothers and fathers (♂) + ♀♀ + ♂♂	Paternity of (♂)	Genetic data from the putative father (♂)
Two-Sex Maternity	Multiple genetic mothers and fathers (♀) + ♀♀ + ♂♂	Maternity of (♀)	Genetic data from the putative mother (♀)
Two-Sex Parentage	Multiple genetic mothers and fathers (♂ + ♀) + ♀♀ + ♂♂	Parentage of (♂ + ♀) Paternity of (♂), Maternity of (♀)	Genetic data from the putative parents (♂ + ♀)
Single-Sex Multiple Mating			
Single-Sex Paternity	Single genetic mother and multiple genetic fathers (♂) + ♀ + ♂♂	Paternity of (♂)	Genetic data from the putative father (♂) and the single genetic mother ♀
Single-Sex Maternity	Single genetic father and multiple genetic mothers (♀) + ♂ + ♀♀	Maternity of (♀)	Genetic data from the putative mother (♀) and the single genetic father ♂

population-wide estimates of parentage. Therefore, there is a need for models that do not require the sampling of all candidate parents and provide parentage estimates of individuals. Our models provide parentage estimates for individuals in both single-sex and two-sex multiple mating systems. They estimate the proportion of offspring fathered or mothered by a putative parent, but do not provide identification of specific parent–offspring relationships. Our approach provides several advantages. First, the models are particularly useful for analysing large sample sizes, where specific NGIs may be less important than the proportion. Second, not all candidate parents are required. Our models require sampling of only sufficient individuals from the breeding population to estimate allele frequencies. Third, the models can be used with many types of genetic data including, for example, allozymes and microsatellites. Fourth, the subset of genetic markers that have the greatest resolving power specific to each putative parent or parent pair can be identified, thereby increasing the efficiency of the analysis. Fifth, the models are robust to minor violations of their assumptions and we provide solutions for major violations, thus increasing their scope. Finally, the models provide unbiased estimates, given the breeding population allele frequencies and random mating with respect to the marker alleles.

In this work, we develop three models. The first two models are for systems with two-sex multiple mating and therefore assume that the NGIs are from multiple males and multiple females that have each mated with several partners. The first model assumes there is genetic data available from only a putative father or a putative mother (but not both) and estimates the paternity or maternity of the putative parent. The second model assumes that there is genetic data available from a putative parent pair and estimates their parentage as well as their individual paternity or maternity. The third model is for systems with single-sex multiple mating and estimates the paternity or maternity of a putative parent. We demonstrated the use of the models using genetic data from a nest of the bluegill sunfish (*Lepomis macrochirus*). In bluegill the offspring within a nest may be from multiple females that have spawned with multiple males from different life histories (cuckolder and parental). The many thousands of eggs and fry are raised by a single 'parental' male, and parental males vary in their degree of paternity. Despite extensive research in bluegill (e.g. Gross 1991; Philipp & Gross 1994), we have not previously been able to provide a precise calculation of an individual's paternity.

The models

The basic framework for our models is limited genetic information from a putative father and/or putative mother plus a sample of NGIs (e.g. a brood). We wish to determine the proportion of the NGIs that are in fact genetically fathered or mothered by the putative parent or parents without having to sample all possible parents in the population. Note that we do not necessarily desire positive identification of the genetic offspring, but the proportion of offspring fathered or mothered by the putative parent or parents (range = 0–100%). Our models require codominant single-locus allelic data (e.g. microsatellite and allozyme data) from the putative parent or parents and the sample of NGIs. They also require an estimate of the frequency of the putative parent's or parents' alleles within the breeding population. Thus, our models require a genetic sample of the breeding population, but do not require the sampling of all potential parents. An overview of the models is presented in Table 1 and Table 2. All variables are defined in Table 3.

Table 2 Summary of the key assumptions, effects of violations, and possible solutions

Assumption	Effects of violation	Comment
1. Random mating with respect to the marker alleles	Assortative mating overestimates success and disassortative mating underestimates success	Assumption is unlikely to be violated if markers are neutral
2. Breeding population allele frequencies are known precisely	Sampling error may overestimate success	With appropriate sampling the overestimate should be negligible (see Appendix IV)
3. Homogeneous allele frequencies	Heterogeneous frequencies can over- or underestimate success	See Appendix II for correction factors
4. NGIs not produced by the putative parent(s) are in Hardy–Weinberg equilibrium	Increases variance in estimated success, but does not bias it	See Appendix IV for bias and Neff *et al*. 2000a for variance
5. Genetic parents of NGIs are not more related than to sampled individuals used to calculate allele frequencies	Relatedness overestimates success	See Appendix III for correction factors

Table 3 Definitions of the variables in the models

Variables Common to All Models

F_{dad}^{l}	Frequency of the alleles in the putative father's (indicated by the subscript 'dad') genotype at locus l. If the putative father is homozygous for allele a then $F_{dad}^{l} = P_{la}$; otherwise if he is heterozygous for alleles a and b then $F_{dad}^{l} = P_{la} + P_{lb}$.
F_{mom}^{l}	Analogous to F_{dad}^{l}.
L	Number of loci used to generate the multilocus genotypes.
Mat	Maternity of the putative mother expressed as a proportion.
Par	Parentage of the putative mother and father expressed as a proportion.
Pat	Paternity of the putative father expressed as a proportion.
P_{la}	Frequency in the population of allele a at locus l.

Variables Common to the Two-Sex Models

ng_{dad}	The observed proportion of the NGIs that is genetically compatible with the putative father (i.e. the proportion that has at least one allele that is indistinguishable from the father's at each of the L loci).
ng_{mom}	Analogous to ng_{dad}.
NG_{dad}	The expected proportion of the NGIs that is genetically compatible with the putative father by chance. Equivalent to the probability that mating between a random male and female from the breeding population would generate a NGI that is genetically compatible with the putative father.
NG_{mom}	Analogous to NG_{dad}.
NG_{dad}^{mepf}	The probability that a mating between the putative mother and a random male from the breeding population (mother's extra-pair fertilization (*mepf*)) would generate a NGI that is genetically compatible with the putative father.
NG_{mom}^{depf}	Analogous to NG_{dad}^{mepf}.

Variables Common to the Two-Sex Parentage and Single-Sex Paternity or Maternity Models

ng_{pair}	The observed proportion of the NGIs that is genetically compatible with the putative parent pair (i.e. the proportion that has at least one allele that is indistinguishable from the putative mother's and the other allele that is indistinguishable from the putative father's at each of the L loci).
NG_{pair}	The expected proportion of the NGIs that is genetically compatible with the putative parent pair by chance. Equivalent to the probability that a mating between a random male and female from the breeding population would generate a NGI that is genetically compatible with the putative parent pair.
NG_{pair}^{mepf}	The probability that a mating between the putative mother and a random male from the breeding population would generate a NGI that is genetically compatible with the putative parent pair.
NG_{pair}^{depf}	Analogous to NG_{pair}^{mepf}.
S_{dad}^{l}	The number of the putative mother's alleles that are identical to either of the putative father's alleles at locus l.
S_{mom}^{l}	The number of the putative father's alleles that are identical to either of the putative mother's alleles at locus l.

The derivation of the first model is presented below and the derivation of the other two, although of similar format, is more complex and these are presented in Appendix I. The derivation of the models follow an approach similar to that of Westneat *et al.* (1987), Wrege & Emlen (1987) and Philipp & Gross (1994).

Two-sex multiple mating

Two-sex paternity or maternity. This model is used to estimate the paternity of a putative father, or the maternity of a putative mother, when there is two-sex multiple mating. The formulas require genetic data from only the putative father, or the putative mother, but not both.

Assume a breeding population in which multiple males and multiple females may contribute their genes to a sample of NGIs. If there is Mendelian inheritance then each NGI will receive one allele from its father and one from its mother at each locus. However, the alleles may also be found in other adults and therefore the presence of a shared allele does not itself provide complete evidence of paternity or maternity. Let the paternity of a putative father, expressed as a proportion, be *Pat* and similarly the maternity of a putative mother be *Mat*. The expected proportion of the NGIs that are genetically compatible with the putative parent (ng_{dad} or ng_{mom}) is calculated from the proportion produced by the putative parent (*Pat* or *Mat*), the proportion produced by other parents (1 – *Pat* or 1 – *Mat*), and the proportion of these NGIs that are expected to be genetically compatible with the putative parent by chance (NG_{dad} or NG_{mom}):

$$ng_{dad} = Pat + (1 - Pat) \cdot NG_{dad};$$

$$ng_{mom} = Mat + (1 - Mat) \cdot NG_{mom};$$

These equations can be solved for *Pat* and *Mat*, providing formulas to calculate the paternity of the putative father or the maternity of the putative mother:

$$Pat = \frac{ng_{dad} - NG_{dad}}{1 - NG_{dad}}; \tag{1}$$

$$Mat = \frac{ng_{mom} - NG_{mom}}{1 - NG_{mom}}. \tag{2}$$

In these two formulas, NG_{dad} and NG_{mom} are calculated from the breeding population allele frequencies under the assumption that the loci are independent (i.e. unlinked):

$$NG_{dad} = \prod_{l=1}^{L}(2 \cdot F_{dad}^{l} \cdot (1 - F_{dad}^{l}) + (F_{dad}^{l})^{2})$$

$$= \prod_{l=1}^{L}(F_{dad}^{l} \cdot (2 - F_{dad}^{l})); \tag{3}$$

$$NG_{mom} = \prod_{l=1}^{L}(F_{mom}^{l} \cdot (2 - F_{mom}^{l})). \tag{4}$$

Here, F_{dad}^{l} is the frequency of the putative father's alleles at locus l (F_{mom}^{l} is analogous; see Table 3 for a detailed explanation of the variables).

Two-sex parentage. This model is used to estimate the parentage of a putative parent pair, as well as their individual paternity and maternity, when there is two-sex multiple mating and when there is genetic data from both the putative father and the putative mother. The same set of data are used for the three calculations. The parentage is the proportion of the NGIs that are produced by the parent pair and it cannot be estimated from the paternity and maternity estimates of the previous models. While this model is more complex, the simultaneous use of genetic data from both parents provides more precise paternity and maternity estimates as well as the parentage estimate. The derivation is provided in Appendix I.

The parentage of the putative parent pair (*Par*), the paternity of the putative father (*Pat*) and the maternity of the putative mother (*Mat*) are estimated in this model by the following formulas:

$$Par = \frac{1}{D} \cdot \begin{pmatrix} NG_{pair}^{depf} \cdot (NG_{dad}^{mepf} \cdot (ng_{mom} - NG_{mom}) + \\ NG_{dad} \cdot (1 - ng_{mom}) - ng_{dad} \cdot (1 - NG_{mom})) + \\ NG_{pair}^{mepf} \cdot (NG_{mom}^{depf} \cdot (ng_{dad} - NG_{dad}) + \\ NG_{mom} \cdot (1 - ng_{dad}) - ng_{mom} \cdot (1 - NG_{dad})) + \\ (1 - NG_{mom}^{depf}) \cdot (ng_{dad} \cdot NG_{pair} - ng_{pair} \cdot NG_{dad}) + \\ (ng_{pair} - NG_{pair}) \cdot (1 - NG_{dad}^{mepf} \cdot NG_{mom}^{depf}) + \\ (1 - NG_{dad}^{mepf}) \cdot (ng_{mom} \cdot NG_{pair} - ng_{pair} \cdot NG_{mom}) \end{pmatrix}; \tag{5}$$

$$Pat = \frac{1}{D} \cdot \begin{pmatrix} (1 - NG_{mom}^{depf}) \cdot (NG_{dad}^{mepf} \cdot (ng_{pair} - NG_{pair}) - \\ NG_{pair}^{mepf} \cdot (ng_{dad} - NG_{dad}) + ng_{dad} \cdot NG_{pair} - \\ ng_{pair} \cdot NG_{dad}) - \\ (1 - NG_{pair}^{depf}) \cdot (NG_{dad}^{mepf} \cdot (ng_{mom} - NG_{mom}) + \\ NG_{dad} \cdot (1 - ng_{mom}) - ng_{dad} \cdot (1 - NG_{mom})) \end{pmatrix}; \tag{6}$$

$$Mat = \frac{1}{D} \cdot \begin{pmatrix} (1 - NG_{dad}^{mepf}) \cdot (NG_{mom}^{depf} \cdot (ng_{pair} - NG_{pair}) - \\ NG_{pair}^{depf} \cdot (ng_{mom} - NG_{mom}) + ng_{mom} \cdot NG_{pair} - \\ ng_{pair} \cdot NG_{mom}) - \\ (1 - NG_{pair}^{mepf}) \cdot (NG_{mom}^{depf} \cdot (ng_{dad} - NG_{dad}) + \\ NG_{mom} \cdot (1 - ng_{dad}) - ng_{mom} \cdot (1 - NG_{dad})) \end{pmatrix}; \tag{7}$$

where

$$D = \begin{pmatrix} (1-NG_{dad}^{mepf}) \cdot (1-NG_{mom}^{depf}) \cdot NG_{pair} - \\ (1-NG_{dad}^{mepf}) \cdot (NG_{mom} \cdot (1-NG_{pair}^{depf}) + NG_{pair}^{depf}) - \\ (1-NG_{mom}^{depf}) \cdot (NG_{dad} \cdot (1-NG_{pair}^{mepf}) + NG_{pair}^{mepf}) - \\ NG_{dad}^{mepf} \cdot NG_{mom}^{depf} + 1 \end{pmatrix}. \tag{8}$$

Single-sex multiple mating

Single-sex paternity or maternity. These formulas are used, when there is single-sex multiple mating, to estimate the paternity of a putative father (the mother's maternity is 1) or the maternity of a putative mother (the father's paternity is 1). They require genetic data from both parents.

The paternity of the putative father (*Pat*), or the maternity of the putative mother (*Mat*), is estimated from the following two formulas (see Appendix I):

$$Pat = \frac{ng_{pair} - NG_{pair}^{mepf}}{1 - NG_{pair}^{mepf}}, \tag{9}$$

$$Mat = \frac{ng_{pair} - NG_{pair}^{depf}}{1 - NG_{pair}^{depf}}. \tag{10}$$

Assumptions, violations and solutions

Our models make five assumptions (Table 2). First, they assume that there is random mating with respect to the marker alleles at each locus. Assortative mating with respect to the marker alleles leads to an overestimation of success, because the number of NGIs that are genetically compatible with the putative parent will be higher than expected. Conversely, disassortative mating leads to an underestimation of success. As genetic markers are generally believed to be neutral, marker-based mating biases should be rare, and therefore we have not developed correction factors.

Second, the models require population allele frequencies. It can be shown that sampling error for these allele frequencies can introduce small biases into estimates from the models. For example, the expected value of *Pat*, in the *Two-Sex Paternity* model, may be less than its true value, thereby underestimating the putative father's paternity. It is therefore useful to minimize sampling error by using appropriate sample sizes (see Appendix IV; Zar 1999). Generally, the bias decreases with greater sampling of the breeding population and increasing resolving power of the loci specific to the putative parent or parents. With appropriate sample sizes, this bias should be small and inconsequential.

Third, the models assume that allele frequencies are the same among all life history types within the breeding population (e.g. males vs. females; subsets within males). For instance, if the frequency of a putative father's alleles is estimated from only breeding males, but the frequency of his alleles is higher in the overall breeding population (males and females), then his paternity will be overestimated. This is because a greater proportion of the parents will carry an allele that is shared with the putative father than is expected, and therefore a greater proportion of the NGIs will be genetically compatible with him. Conversely, if the frequency of the putative father's alleles is lower in the overall breeding population than the sampled breeding males, then his paternity will be underestimated. This bias can be high. For example, if the putative father's alleles are twice as frequent in the overall breeding population, then his paternity can be overestimated by more than 10% (data not shown). In Appendix II we therefore develop correction factors that remove this bias.

Fourth, the models assume that within the sample of NGIs, offspring not produced by the putative parent or parents conform to the Hardy–Weinberg genotype proportions of the population. If only a select group of adults contribute to the NGIs, or if the sample of NGIs is small (e.g. a small brood), then Hardy–Weinberg genotype proportions are unlikely to be realized. We show in Appendix IV, however, that no bias is introduced into the estimate. For example, we show that in the *Two-Sex Paternity* model the expected value of NG_{dad} is independent of the number of mothers and fathers that contribute to the sample of NGIs. The significance of violating this assumption appears only in the variance of the estimate (Neff *et al.* 2000a). While the variance is increased, the estimate itself is unbiased.

Fifth, the models assume that the parents of NGIs are not more related than the sample from which the allele frequencies are based. If the parents are genetically related and the allele frequencies have been estimated from a sample of unrelated individuals, the models will overestimate the true success of the putative parent or parents. In Appendix III we develop correction factors that remove this bias.

Biological example

Bluegill sunfish (*Lepomis macrochirus*) are found in freshwater lakes throughout much of central and eastern North America (Lee *et al.* 1980). They spawn in colonies with as many as 300 nests (Cargnelli & Gross 1996). Male life histories have a discrete polymorphism termed 'parental' and 'cuckolder' (Gross 1982, 1996). In Lake Opinicon, Ontario, parental males mature at 7 years of age and compete to construct a nest in a colony. Cuckolder males mature precociously at 2 years of age and steal fertilizations by intruding while females are spawning in the nests of parental males. Cuckolders do not build their own nests or guard their offspring.

Table 4 Data from a biological example: paternity in a parental male bluegill sunfish (*Lepomis macrochirus*). The parental male's genotype, allele frequencies, NG_{dad} and the cumulative product of NG_{dad} are presented for six microsatellite loci. The loci are arranged in ascending order of NG_{dad}

Locus	Genotype	Allele frequency	NG_{dad}*	ΠNG_{dad}†
Lma102	88/88	0.33	0.551	0.551
Lma120	227/247	0.33/0.01	0.563	0.310
Lma87	118/118	0.42	0.658	0.204
Lma20	111/111	0.58	0.823	0.168
Lma117	194/206	0.45/0.21	0.884	0.148
Lma121	190/190	0.66	0.884	0.131

*Calculated from eqn 3.
†As an example, the cumulative product for Lma87 is the products of NG_{dad} for Lma102, Lma120 and Lma87: $0.551 \times 0.563 \times 0.658 = 0.204$.

Only the resident parental male provides care for the developing eggs and fry in his nest. Spawning often involves multiple males and females in a single nest and results in several thousands of embryos of mixed parentage being raised by a single parental male. The models developed here can be used to determine the paternity of an individual parental male within the fry (NGIs) in his nest.

We captured a parental male and a sample of the fry from his nest in a natural colony in Lake Opinicon (June 1996). To estimate the breeding population allele frequencies, we collected postspawning females ($N = 44$), parental males ($N = 106$) and cuckolder males ($N = 82$) free-swimming in the vicinity of the colony. Because of the mating dynamics and as we did not have a putative mother, we applied the *Two-Sex Paternity* model, which enables the estimation of paternity of a single male and allows for multiple mating by both sexes.

We calculated the allele frequencies of six microsatellites in the population sample of 232 adults ($N = 44 + 106 + 82 = 232$; microsatellite techniques are in Neff *et al*. 2000b). We obtained genotypes from the putative father (i.e. the parental male) for the six loci, and calculated the value of NG_{dad} for each locus from his genotype and the estimated population allele frequencies. We chose the three loci with the lowest values of NG_{dad} (i.e. the greatest resolving power) and obtained their genotypes from 46 fry. The data were entered into the model (eqn 1) to estimate the putative father's paternity.

As the allele frequencies did not differ among parental males, cuckolder males and females, the frequency data from the 232 adults were combined. Table 4 presents the putative father's genotypes and the corresponding NG_{dad} values. Table 5 presents the three microsatellite loci chosen from Table 4 and the genotypes of the 46 offspring. Table 6 presents the paternity estimates based on

Table 5 Summary of the genotypes of 46 offspring at three microsatellite loci: Lma102, Lma120 and Lma87. The parental male's genotype is 88/88, 227/247 and 118/118. *N* refers to the number of offspring with the multilocus genotype. Bold numbers indicate fry that are genetically incompatible with the parental male. The proportion of the offspring that are compatible with the parental male at each locus and all three loci is indicated

Lma102	Lma120	Lma87	N
88/98	217/227	118/128	3
88/98	217/227	118/152	1
88/98	217/247	118/128	1
88/98	217/247	118/152	5
88/98	227/227	118/128	1
88/98	227/227	118/152	3
88/98	227/231	**128/152**	**1**
88/98	227/247	118/128	4
88/98	227/247	118/152	2
88/102	217/227	118/128	3
88/102	217/227	118/152	2
88/102	217/247	118/128	5
88/102	217/247	118/152	1
88/102	227/227	118/128	2
88/102	227/227	118/152	3
88/102	227/247	118/128	1
88/102	227/247	118/152	3
98/98	**211/231**	118/128	**1**
98/98	**227/231**	**128/128**	**1**
98/98	**231/245**	118/128	**1**
98/102	**211/231**	118/128	**1**
98/102	**211/231**	**128/152**	**1**
41/46	42/46	43/46	40/46
(89.1%)	(91.3%)	(93.5%)	(87.0%)

each of the three loci individually, each pair of loci and all three loci together. They range from 75.8 to 84.2%. Thus, we might infer that the parental male fertilized ≈ 76–84% of the NGIs within his nest and under his care (but see Neff *et al*. 2000a).

Discussion

The models

We have presented new models for calculating parentage using genetic markers. The models are based on knowledge of the genotypes of putative parents, genotypes of NGIs and allele frequencies within the breeding population. They are especially useful when genetic data are limited, such as when not all candidate parents are available or when it is not possible to exclude all but one parent or parent pair. They can be used when there is multiple mating by either one sex or both sexes. Thus, complex mating systems can be addressed. The models can utilize loci with numerous alleles and enable the identification of the subset of genetic markers that have

Loci	NG_{dad}	ng_{dad} (%)*	Pat (%)
Single			
Lma102	0.551	89.1 (41/46)	75.8
Lma120	0.563	91.3 (42/46)	80.1
Lma87	0.658	93.5 (43/46)	80.9
Paired			
Lma102, Lma120	0.310	89.1 (41/46)	84.2
Lma102, Lma87	0.362	87.0 (40/46)	79.6
Lma120, Lma87	0.370	87.0 (40/46)	79.3
All			
Lma102, Lma120, Lma87	0.204	87.0 (40/46)	83.6
Range	0.204–0.658	87.0–93.5	75.8–84.2

Table 6 The paternity results for the parental male bluegill. NG_{dad} and ng_{dad} are defined in Table 3. *Pat* is the paternity estimate for the parental male based on single loci, paired loci, and all three loci collectively. The range in values is also indicated

*ng_{dad} is calculated from Table 5.

the greatest resolving power specific to each putative parent, thereby increasing the efficiency of the analysis. The models are robust, at least to moderate violations of their assumptions. We have developed corrections for significant violations, including when allele frequencies are not homogeneous among groups of individuals and when parents are genetically related. The models provide unbiased estimates given the breeding population allele frequencies and random mating with respect to the marker alleles.

Previously published methods for parentage inference in natural populations have focused on assigning progeny among a set of candidate parents. These methods include exclusion (e.g. Chakraborty *et al*. 1988), categorical allocation of offspring to the most-likely parent or parent pair (e.g. Meagher & Thompson 1986, 1987; Meagher 1986) and fractional allocation of offspring among all nonexcluded parents (e.g. Devlin *et al*. 1988; Smouse & Meagher 1994). Exclusion methods can provide accurate parentage inference, but may require a large number of loci to exclude all but the true parents (Chakraborty *et al*. 1988). Categorical allocation does not require exclusion of all but the true parents and has the potential benefit of identifying the single most-likely parent or parent pair, and has become common in parentage studies (e.g. Coltman *et al*. 1998; Prodöhl *et al*. 1998). A limitation of these models is the implicit assumption that the entire pool of candidate parents has been sampled. Marshall *et al*. (1998) have shown that incomplete sampling of the pool of candidate parents can reduce the accuracy of the parentage inference. In many cases, complete sampling will not be feasible. By contrast, the models developed in this paper do not require the sampling of the entire pool of candidate parents and will be particularly useful when not all of the parents are available.

Westneat *et al*. (1987) and Wrege & Emlen (1987) have developed methods that do not require the sampling of all candidate parents and provide population-wide estimates of the rates of extrapair fertilizations (EPFs), intraspecific egg parasitism (egg dumping) and quasi-parasitism. They define an EPF as an offspring that belongs to a resident female, but not to the resident male. Intraspecific egg parasitism is an offspring that belongs to a female other than the resident female. Quasi-parasitism is a special case of intraspecific egg parasitism where the resident male fertilizes an egg not belonging to the resident female. Our models can be used to estimate rates of these mating types for individuals. From the *Two-Sex Parentage* model, the rate of EPFs within a female's nest is estimated as the putative mother's maternity less the putative parent pair's parentage. The rate of intraspecific egg parasitism is estimated as one minus the putative father's paternity minus the putative mother's maternity plus their parentage. The rate of quasi-parasitism is estimated as the putative father's paternity minus the putative parent pair's parentage. If there is no egg parasitism then there is multiple mating by only one sex, and the *Single-Sex Paternity* model can be used to provide a more precise estimate of the frequency of EPFs.

It is important to distinguish between the breeding population, as we define it, and the population at large. For our models, the breeding population refers specifically to only the individuals that contribute their genes to a sample of NGIs. These individuals may represent a proportion of the population at large or even a proportion of all breeding individuals (i.e. not all sexually mature individuals may contribute their genes to a given sample of NGIs). The models require allele frequencies for the breeding population specific to a sample of NGIs. Where it is possible to determine the group of breeding parents, such as a communally breeding bird, sampling the dominant breeders can provide accurate allele frequencies without having to sample from the entire population, and will provide the most precise parentage estimates. However, if there are multiple independent samples of NGIs (e.g. the broods from different nests), then a sample from

the entire breeding population may be easier to obtain than samples from each portion of the breeding population specific to the different nests. Provided the breeding individuals specific to each nest are a random sample (with respect to marker genotypes) of the entire breeding population, then a sample from the latter will provide accurate allele frequencies and unbiased parentage estimates (see Appendix IV). Furthermore, if the breeding population is a random sample (with respect to marker genotypes) of the population at large, then any sample can provide the necessary allele frequencies. Such a sample is often readily available. For example, the nest-guarding parental males collected along with the offspring from a colony of bluegill sunfish can provide accurate allele frequencies.

In many mating systems, the breeding individuals specific to a sample of NGIs may be genetically related (e.g. Packer *et al.* 1991; Davies 1992). It is therefore desirable to develop models that will allow for the genetic relatedness of parents. In Appendix III, we quantified the effects of relatedness on the parentage estimates and derived parameters that allow for relatedness. For instance, when the breeding individuals genetically contributing to a sample of NGIs are more related than the sample used to calculate allele frequencies, the models will overestimate success. In such cases, the modified parameters account for relatedness and provide unbiased parentage inference. This should increase the scope of application of the models. It should be noted that when the breeding individuals contributing to a sample of NGIs are not more related than the sample used to calculate the allele frequencies, the correction parameters are not necessary. As an example, if the breeding population as a whole is equally related, or if the specific subset of breeding individuals are sampled, then the relatedness is accounted for in the allele frequency estimates and therefore the relatedness parameters are unnecessary (also see Appendix III).

The biological example

The bluegill sunfish mating system is amongst the more complex known, because multiple fathers (including different life histories) and multiple mothers contribute to a single nest of NGIs. We used the *Two-Sex Paternity* model as there was multiple mating within both sexes and because we did not have a putative mother. The three loci provided consistent estimates for the paternity of the parental male bluegill (75.8–84.2%). The estimate based on all three loci collectively (83.6%) should be the most precise as it is based on the greatest amount of information (see Neff *et al.* 2000a). Thus, some 16–17% of the progeny that the parental male was rearing were fathered by other males, such as cuckolders. If we had another putative father (e.g. a cuckolder male) then we

could also obtain estimates of his individual paternity within the nest. If we had a putative mother (e.g. a female observed spawning with the parental male) then we could apply the *Two-Sex Parentage* model to provide a more precise estimate of the parental male's paternity as well as estimate the putative mother's maternity and their parentage.

In this example, the *Two-Sex Paternity* model modified the parental male's success only marginally from the total proportion of NGIs that were compatible ([87–83.6]/87 = 4%). However, this represents a significant increase of greater than 26% in the estimate of cuckolder success ([16.4–13]/13 = 26.2%). Further, these differences increase as the putative father's success decreases and can be quite significant (see Neff *et al.* 2000a). Finally, estimating paternity strictly as the proportion of the sample that is compatible with the putative father will lead to a systematic bias (overestimate) in success. This can be particularly concerning when, for example, calculating the fitness of alternative life histories. Therefore, the more sophisticated estimates from our models should be used to provide unbiased parentage inference.

The models developed in this work provide estimates of individual parentage in complex mating systems with limited genetic information. These estimates have not previously been attainable.

References

Avise JC (1994) Molecular markers. *Natural History and Evolution*. Chapman & Hall, New York, NY.

Cargnelli L, Gross MR (1996) The temporal dimension in fish recruitment: birth date, body size, and size-dependent survival in a sunfish (bluegill: *Lepomis macrochirus*). *Canadian Journal of Fisheries and Aquatic Sciences*, **53**, 360–367.

Chakraborty R, Meagher TR, Smouse PE (1988) Parentage ana-lysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics*, **118**, 527–536.

Coltman DW, Bowen WD, Wright JM (1998) Male mating success in an aquatically mating pinniped, the harbour seal (*Phoca vitulina*), assessed by microsatellite DNA markers. *Molecular Ecology*, **7**, 627–638.

Davies NB (1992) *Dunnock Behaviour and Social Evolution*. Oxford University Press, Oxford, N.Y.

Devlin D, Roeder K, Ellstrand NC (1988) Fractional paternity assignment: Theoretical development and comparison to other methods. *Theoretical and Applied Genetics*, **76**, 369–380.

Ellstrand NC (1984) Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. *American Naturalist*, **123**, 819–828.

Gross MR (1982) Sneakers, satellites and parentals: polymorphic mating strategies in North American sunfishes. *Zeitschrift Fur Tierpsychologie*, **60**, 1–26.

Gross MR (1991) Evolution of alternative reproductive strategies: frequency-dependent sexual selection in male bluegill sunfish. *Philosophical Transactions of the Royal Society of London B*, **332**, 59–66.

Gross MR (1996) Alternative reproductive strategies and tactics: diversity within sexes. *Trends in Ecology and Evolution*, **11**, 92–98.

Jarne P, Lagoda JL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, **8**, 285–288.

Keane B, Dittus WPJ, Melnick DJ (1997) Paternity assignment in wild groups of toque macaques *Macaca sinica* at Polonnaruwa Sri Lanka using molecular markers. *Molecular Ecology*, **6**, 267–282.

Lee DS, Gilbert CR, Hocutt CH, Jenkins RE, McAllister DE, Stauffer JR Jr (1980) *Atlas of North American Freshwater Fishes*. North Carolina State Museum of Natural History, NC, USA.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.

Meagher TR (1986) Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely male parents. *American Naturalist*, **128**, 199–215.

Meagher TR, Thompson E (1986) The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology*, **29**, 87–106.

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

Morin PA, Wallis J, Moore JJ, Woodruff DS (1994) Paternity exclusion in a community of wild chimpanzees using hyper-variable simple sequence repeats. *Molecular Ecology*, **3**, 469–478.

Neff BD, Repka J, Gross MR (2000a) Statistical confidence in parentage analysis with incomplete sampling: How many loci and offspring are needed? *Molecular Ecology*, **9**, 529–539.

Neff BD, Fu P, Gross MR (2000b) Microsatellite multiplexing in fish. *Transactions of the American Fisheries Society*, **129**, 590–599.

Packer C, Gilbert DA, Pusey AE, O'Brien SJ (1991) A molecular genetic analysis of kinship and cooperation in African lions. *Nature*, **351**, 562–565.

Petrie M, Kempenaers B (1998) Extra-pair paternity in birds: explaining variation between species and populations. *Trends in Ecology and Evolution*, **13**, 52–58.

Philipp DP, Gross MR (1994) Genetic evidence of cuckoldry in bluegill *Lepomis macrochirus*. *Molecular Ecology*, **3**, 563–569.

Prodöhl PA, Loughry WJ, McDonough CM, Nelson WS, Thompson EA, Avise JC (1998) Genetic maternity and paternity in a local population of Armadillos assessed by microsatellite DNA markers and field data. *American Naturalist*, **151**, 7–19.

Reynolds JD (1996) Animal breeding systems. *Trends in Ecology and Evolution*, **11**, 68–72.

Smouse PE, Meagher TR (1994) Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (*Liliaceae*). *Genetics*, **136**, 313–322.

Westneat DF, Frederick PC, Wiley RH (1987) The use of genetic markers to estimate the frequency of successful alternative reproductive tactics. *Behavioural Ecology and Sociobiology*, **21**, 35–45.

Wrege PH, Emlen ST (1987) Biochemcial determination of parental uncertainty in white-fronted bee-eaters. *Behavioural Ecology and Sociobiology*, **20**, 153–160.

Zar JH (1999) *Biostatistical Analysis*, 4th edn. Prentice Hall, Inc., Simon & Schuster, Upper Saddle River, NJ.

Appendix I

Derivations of the *Two-Sex Parentage* and *Single-Sex Paternity or Maternity* models. All variables are defined in Table 3.

Two-Sex Parentage

Assume a breeding population in which multiple males and multiple females may be contributing their genes to a sample of next-generation individuals (NGIs), e.g. a brood. We are interested in estimating the proportion of the sample of NGIs that were produced by a particular male and female pair, the putative parents. We therefore modify the first model to include the genetic contribution of both putative parents to calculate their 'parentage'. There are four steps. The first three steps develop formulas for calculating the expected proportion of the sample of NGIs that are genetically compatible with the putative father, the putative mother and both putative parents, respectively. The fourth step combines these formulas using linear algebra to develop formulas for estimating the putative parents' parentage as well as their individual paternity and maternity. We provide a more detailed development of step one. Steps two and three follow an analogous approach.

1. Considering a sample of NGIs, there are three possible matings that could produce NGIs that are compatible with the putative father. First, all of the NGIs fertilized by the putative father will be compatible with him (i.e. his paternity, *Pat*). This includes matings between the putative father and the putative mother, as well as with other females. Second, a proportion of NGIs produced from matings between individuals other than the putative parents will be compatible by chance. This proportion will depend on the chance that a shared allele (at each locus) is passed to the offspring by either the true mother or father. The less common the putative father's genotype is, the smaller this chance and proportion will be. Third, a proportion of NGIs produced from matings between the putative mother and males other than the putative father will also be compatible by chance. This proportion depends on the chance that the true father or the putative mother passes a compatible allele to the offspring. Including these three mating types, the proportion of the NGIs expected to be compatible with the putative father can be calculated from:

$$ng_{dad} = Pat + (1 - Pat - Mat + Par) \cdot NG_{dad} +$$
$$(Mat - Par) \cdot NG_{dad}^{mepf}; \qquad \text{(A1.1)}$$

Briefly, the expected proportion of the NGIs that are genetically compatible with the putative father (ng_{dad}) is equal to his paternity (*Pat*) plus the proportion of NGIs produced by individuals other than the putative parents $(1 - Pat - Mat + Par)$ multiplied by the proportion of these NGIs expected to be compatible with the putative father by chance (NG_{dad}) plus the proportion of NGIs produced by the putative mother and males other than the putative father ($Mat - Par$) multiplied by the proportion of these NGIs expected to be compatible with the putative father by chance (NG_{dad}^{mepf}).

NG_{dad} has been calculated earlier (eqn 3). NG_{dad}^{mepf} can be calculated in terms of the putative father's allele frequencies (F_{dad}^l) and the number of alleles that he shares with the putative mother (S_{dad}^l) as follows:

$$NG_{dad}^{mepf} = \prod_{l=1}^{L} (F_{dad}^l + \tfrac{1}{2} \cdot S_{dad}^l - F_{dad}^l \cdot \tfrac{1}{2} \cdot S_{dad}^l). \qquad \text{(A1.2)}$$

2. The expected proportion of the sample of NGIs that are genetically compatible with the putative mother is analogous to eqn A1.1 and can be calculated from:

$$ng_{mom} = Mat + (1 - Pat - Mat + Par) \cdot NG_{mom} +$$
$$(Pat - Par) \cdot NG_{mom}^{depf}; \qquad \text{(A1.3)}$$

where

$$NG_{mom}^{depf} = \prod_{l=1}^{L} (F_{mom}^l + \tfrac{1}{2} \cdot S_{mom}^l - F_{mom}^l \cdot \tfrac{1}{2} \cdot S_{mom}^l). \qquad \text{(A1.4)}$$

3. The expected proportion of the sample of NGIs that are genetically compatible with both of the putative parents is:

$$ng_{pair} = Par + (1 - Pat - Mat + Par) \cdot NG_{pair} +$$
$$(Pat - Par) \cdot NG_{pair}^{depf} + (Mat - Par) \cdot NG_{pair}^{mepf}. \quad \text{(A1.5)}$$

Briefly, the expected proportion of the NGIs that are genetically compatible with both of the putative parents (ng_{pair}) is equal to their parentage (*Par*; all of this proportion is compatible with them) plus the proportion of NGIs produced by individuals other than the putative parents $(1 - Pat - Mat + Par)$ multiplied by the proportion of these NGIs expected to be compatible with the putative parents by chance (NG_{pair}) plus the proportion of NGIs produced by the putative father and females other than the putative mother ($Pat - Par$) multiplied by the proportion of these NGIs expected to be compatible with the putative parents by chance (NG_{pair}^{depf}) plus the proportion of NGIs produced by the putative mother and males other than the putative father ($Mat - Par$) multiplied by the proportion of these NGIs expected to be compatible with the putative parents by chance (NG_{pair}^{mepf}).

NG_{pair} can be calculated from the allele frequencies of the putative parents (P_l) as follows:

$$NG_{pair} = \prod_{l=1}^{L}\sum_{j=1}^{2}\sum_{k=1}^{2}(c_{jk} \cdot b_{jk} \cdot P_{lj} \cdot P_{lk}); \qquad (A1.6)$$

where b_{jk} equals 1 when the father's allele j is indistinguishable from the mother's allele k; otherwise b_{jk} equals 2; and c_{jk} equals 1 when the genotype generated from the union of the father's allele j with the mother's allele k is unique from all genotypes generated from the union of the father's allele x with the mother's allele y for $0 < x < j$ and $0 < y < k$; otherwise c_{jk} equals 0. A mating between the putative father and a female other than the putative mother produces a NGI that is compatible with both putative parents (NG_{pair}^{depf}) when the female shares an allele (at each locus) with the putative mother and contributes this allele to the NGI. A compatible NGI is also produced when the putative father shares one allele with the putative mother and contributes it to the NGI and his second allele (which, owing to the independence of probabilities, cannot be shared with the putative mother) is contributed by the female that he mates with. In this case, one allele is compatible with the putative mother (from the putative father) and the second allele is compatible with the putative father (from the female). Incorporating both of these possibilities, the probability that a mating between the putative father and a random female from the breeding population produces an offspring that is compatible with the putative parents can be calculated from:

$$NG_{pair}^{depf} = \prod_{l=1}^{L}(\tfrac{1}{2} \cdot P_{lu} + F_{mom}^{l}); \qquad (A1.7)$$

here P_{lu} is the frequency of the putative father's unshared allele when he is heterozygous and shares exactly one allele with the putative mother at locus l; otherwise P_{lu} equals 0.

NG_{pair}^{mepf} is analogous to NG_{pair}^{depf} and can be calculated from:

$$NG_{pair}^{mepf} = \prod_{l=1}^{L}(\tfrac{1}{2} \cdot P_{lu} + F_{dad}^{l}); \qquad (A1.8)$$

here P_{lu} is the frequency of the putative mother's unshared allele when she is heterozygous and shares exactly one allele with the putative father at locus l; otherwise P_{lu} equals 0.

4. From eqns A1.1, A1.3 and A1.5 developed in the first three steps, we can obtain solutions for Par, Pat and Mat that include the genetic contribution of both putative parents. As we have three linear eqns (A1.1, A1.3, A1.5), we can solve for the three unknowns using linear algebra (see eqns 5, 6 and 7).

Single-Sex Paternity or Maternity

This model follows directly from the *Two-Sex Parentage* model with the assumption that there is either a single genetic mother, or a single genetic father. If there is only a single genetic mother then her maternity is 1 ($Mat = 1$) and the paternity of the putative father is equivalent to his parentage with the putative mother ($Par = Pat$). Substituting $Mat = 1$ and $Par = Pat$ into eqn A1.5 we get an equation with a single unknown (Pat) that can be solved. Briefly, a proportion of the NGIs are compatible with the putative father and the single genetic mother by chance (NG_{pair}^{mepf}) and the putative father's paternity is equal to the fraction of the remainder ($1 - NG_{pair}^{mepf}$) that are also compatible ($ng_{pair} - NG_{pair}^{mepf}$) (see eqn 9). Similarly, if there is only a single genetic father then $Pat = 1$ and $Par = Mat$ and from eqn A1.5 we get an equation with a single unknown (Mat) that can be solved for the maternity of the putative mother (see eqn 10).

Appendix II

Violation of assumption no. 3 from Table 2.

In this appendix we present generalizations of the earlier equations that can be applied in situations where allele frequency distributions differ among groups of individuals within the breeding population.

When there are two or more male or female groups within the breeding population that have different allele frequency distributions, corrected allele frequencies that represent a weighted average of the allele frequencies within each group should be used. These frequencies can be calculated from:

$$P_{la}^{males} = \sum_{i=1}^{MG} mg_i \cdot P_{la}^{i}; \qquad (A2.1)$$

$$P_{la}^{females} = \sum_{j=1}^{FG} fg_j \cdot P_{la}^{j}; \qquad (A2.2)$$

where P_{la}^{males} is the corrected allele frequency of allele a at locus l within the male breeding population; $P_{la}^{females}$ is analogous; P_{la}^{i} is the frequency of allele a at locus l within male group i; P_{la}^{j} is the frequency of allele a at locus l within female group j; mg_i is the proportion of the next-generation individuals (NGIs) produced by male group i individuals ($\sum mg_i = 1$); fg_j is the proportion of the NGIs produced by female group j individuals ($\sum fg_j = 1$); MG is the number of male groups in the breeding population with different allele frequency distributions; and FG is the number of female groups in the breeding population with different allele frequency distributions.

To calculate NG_{dad} (NG_{mom} is analogous) use the following formula:

$$NG_{dad} = \prod_{l=1}^{L}(F_{dad}^{l,males}+F_{dad}^{l,females}-F_{dad}^{l,males}\cdot F_{dad}^{l,females}); \qquad (A2.3)$$

where $F_{dad}^{l,males}$ is the frequency of the alleles in the putative father's genotype at locus l within the male breeding population. If the putative father is homozygous for allele a then $F_{dad}^{l,males} = P_{la}^{males}$; otherwise $F_{dad}^{l,males} = P_{la}^{males} + P_{la}^{males}$; and $F_{dad}^{l,females}$ is analogous to $F_{dad}^{l,males}$.

To calculate NG_{pair} use the following formula:

$$NG_{pair} = \prod_{l=1}^{L}\sum_{j=1}^{2}\sum_{k=1}^{2}(c_{jk}\cdot b_{jk}\cdot(P_{lj}^{males}\cdot P_{lk}^{females}+P_{lj}^{females}\cdot P_{lk}^{males}));$$

$$(A2.4)$$

where b_{jk} equals $1/2$ when the father's allele j is indistinguishable from the mother's allele k; otherwise b_{jk} equals 1; and c_{jk} equals 1 when the genotype generated from the union of the father's allele j with the mother's allele k is unique from all genotypes generated from the union of the father's allele x with the mother's allele y for $0 < x < j$ and $0 < y < k$; otherwise c_{jk} equals 0.

To calculate NG_{dad}^{mepf} (NG_{mom}^{depf} is analogous) use the following formula:

$$NG_{dad}^{mepf} = \prod_{l=1}^{L}(F_{dad}^{l,males}+\tfrac{1}{2}\cdot S_{dad}^{l}-F_{dad}^{l,males}\cdot\tfrac{1}{2}\cdot S_{dad}^{l}). \qquad (A2.5)$$

To calculate NG_{pair}^{mepf} (NG_{pair}^{depf} is analogous) use the following formula:

$$NG_{pair}^{mepf} = \prod_{l=1}^{L}(\tfrac{1}{2}\cdot P_{lu}^{males}+F_{dad}^{l,males}); \qquad (A2.6)$$

where P_{lu}^{males} is the frequency of the putative mother's unshared allele when she is heterozygous and shares exactly one allele with the putative father at locus l; otherwise P_{lu}^{males} equals 0.

Appendix III

Violation of assumption no. 5 from Table 2.

In this appendix we present generalizations of the earlier equations that can be applied in situations where the genetic parents of a sample of next-generation individuals (NGIs) are related (i.e. kin) and allele frequencies are based on a random sample of individuals.

First, define the degree of relatedness (R) ranging from 0 to 1, so that, for example, $R = 0$ when two individuals are unrelated, $R = 1/2$ when two individuals are full-sibs or parent and offspring and $R = 1$ when two individuals are genetically identical. When the mothers or the fathers contributing to a sample of NGIs are related ($R > 0$), but

the allele frequencies are based on a random sample of the population, the following modified formulas should be used. These correction factors do not consider incestuous mating (e.g. when mothers are related to their mates). Therefore, they assume that the putative father is genetically unrelated to the female breeding parents (including the putative mother) and similarly that the putative mother is genetically unrelated to the male breeding parents. Note that if the allele frequencies are based on the breeding population specific to the sample of NGIs, then these correction factors are not required (see the Discussion).

To calculate the analogues of NG_{dad} and NG_{mom} in situations where the putative parents are related to the other breeding parents, use the following formulas:

$$^{R}NG_{dad} = \prod_{l=1}^{L}(R_{dad}^{males}+(1-R_{dad}^{males})\cdot F_{dad}^{l}+F_{dad}^{l}-$$
$$(R_{dad}^{males}+(1-R_{dad}^{males})\cdot F_{dad}^{l})\cdot F_{dad}^{l})$$
$$= \prod_{l=1}^{L}(R_{dad}^{males}+(1-R_{dad}^{males})\cdot NG_{dad}); \qquad (A3.1)$$

$$^{R}NG_{mom} = \prod_{l=1}^{L}(R_{mom}^{females}+(1-R_{mom}^{females})\cdot NG_{mom}); \qquad (A3.2)$$

where R_{dad}^{males} is the average degree of relatedness (R) between the putative father and each of the other male parents that produced the NGIs; and $R_{mom}^{females}$ is the average degree of relatedness (R) between the putative mother and each of the other female parents that produced the NGIs.

Briefly, eqn A3.1 represents the probability that an allele that is compatible with the putative father is contributed to a NGI by a related breeding male ($R_{dad}^{males} + (1 - R_{dad}^{males})\cdot F_{dad}^{l}$) or an unrelated breeding female (F_{dad}^{l}). As these probabilities are not independent, their product must be subtracted. Simplifying this equation gives the final result.

To calculate the analogue of NG_{pair} use the following formula:

$$^{R}NG_{pair} = \prod_{l=1}^{L}\sum_{j=1}^{2}\sum_{k=1}^{2}c_{jk}\cdot((R_{dad}^{males}\cdot d_{dad}+(1-R_{dad}^{males})\cdot P_{lj})\cdot$$
$$(R_{mom}^{females}\cdot d_{mom}+(1-R_{mom}^{females})\cdot P_{lk})+A\cdot B); \qquad (A3.3)$$

where c_{jk} equals 1 when the genotype generated from the union of the putative father's allele j with the putative mother's allele k is unique from all genotypes generated from the union of the putative father's allele x with the putative mother's allele y for $0 < x < j$ and $0 < y < k$; otherwise c_{jk} equals 0; d_{dad} equals 1 if the putative father is

homozygous; otherwise d_{dad} equals $^1/_2$; d_{mom} is analogous to d_{dad}; and A and B are defined as follows. If the putative father's allele j is equivalent to the putative mother's allele k (at locus l) then $A = 0$ and $B = 0$. If the putative father's allele j is distinct from both of the putative mother's alleles then $A = (1 - R_{mom}^{females}) \cdot P_{lj}$; otherwise $A = R_{mom}^{females} \cdot d_{mom} + (1 - R_{mom}^{females}) \cdot P_{lj}$. Similarly, if the putative mother's allele k is distinct from both of the putative father's alleles then $B = (1 - R_{dad}^{males}) \cdot P_{lk}$; otherwise $B = R_{dad}^{males} \cdot d_{dad} + (1 - R_{dad}^{males}) \cdot P_{lk}$. Briefly, eqn A3.3 represents the probability that the true father contributes an allele equivalent to the putative father's allele j ($R_{dad}^{males} \cdot d_{dad} + (1 - R_{dad}^{males}) \cdot P_{lj}$) and the true mother contributes an allele equivalent to the putative mother's allele k ($R_{mom}^{females} \cdot d_{mom} + (1 - R_{mom}^{females}) \cdot P_{lk}$) or (given that j and k are distinct) the true mother contributes an allele equivalent to the putative father's allele j (A) and the true father contributes an allele equivalent to the putative mother's allele k (B).

To calculate the analogue for NG_{dad}^{mepf} (NG_{mom}^{depf} is analogous) use the following formula:

$$^R NG_{dad}^{mepf} = \prod_{l=1}^{L} (R_{dad}^{males} + (1 - R_{dad}^{males}) \cdot NG_{dad}^{mepf}); \qquad (A3.4)$$

To calculate the analogue for NG_{pair}^{mepf} (NG_{pair}^{depf} is analogous) use the following formula:

$$^R NG_{pair}^{mepf} = \prod_{l=1}^{L} (R_{dad}^{males} + (1 - R_{dad}^{males}) \cdot NG_{pair}^{mepf}). \qquad (A3.5)$$

Appendix IV

A proof that if the population allele frequencies are known and mating is random with respect to the genetic markers, the models provide unbiased estimates of paternity, maternity and parentage, independent of the number of genetic parents contributing to a finite sample of NGIs. The proof is for the *Two-Sex Paternity* model. The proofs for the other models are analogous in format.

From eqn 1, we can calculate the expected paternity for the putative father as:

$$\overline{Pat} = \frac{\overline{ng}_{dad} - N\hat{G}_{dad}}{1 - N\hat{G}_{dad}}, \qquad (A4.1)$$

which depends on the proportion of NGIs that are expected to be compatible with the putative father ($\overline{ng}_{dad}$) and the proportion expected by chance ($N\hat{G}_{dad}$). Here we use the 'hat' to distinguish the estimated value, which is calculated from the allele frequencies sampled. Given these estimated frequencies and the putative father's genotype, $N\hat{G}_{dad}$ is a constant. The proportion $\overline{ng}_{dad}$

depends on the putative father's actual paternity (Pat) and the proportion of NGIs that are not from the putative father ($1 - Pat$), but are compatible by chance (δ):

$$ng_{dad} = Pat + (1 - Pat) \cdot \delta. \qquad (A4.2)$$

The expected value δ is independent of the number of parents that contribute to the NGIs and the number of NGIs sampled. As an example, suppose that M mothers and F fathers contribute equally to the $(1 - Pat)$ sample of NGIs (note that the following proof is analogous if the parents' genetic contributions are unequal). To calculate the expected proportion of these NGIs that are compatible with the putative father we need to define two properties from the binomial theorem. First, the binomial expansion is:

$$\sum_{i=0}^{n} \binom{n}{i} p^i \cdot (1-p)^{n-i} = (p + (1-p))^n = 1. \qquad (A4.3)$$

Second, the expected value of i is:

$$\sum_{i=0}^{n} \binom{n}{i} p^i \cdot (1-p)^{n-i} \cdot i = p \cdot n. \qquad (A4.4)$$

The expected value of δ is equal to the probability that one of the M genetic mothers or F genetic fathers contribute a compatible allele (with the putative father) to a NGI at each of the l loci. This can be calculated from the binomial theorem:

$$\delta = \prod_{l=1}^{L} \sum_{i=0}^{2M} \sum_{j=0}^{2F} \left(\binom{2M}{i} \binom{2F}{j} (F_{dad}^l)^{i+j} \cdot (1 - F_{dad}^l)^{2M+2F-i-j} \cdot \left(\frac{i}{2M} + \frac{j}{2F} - \frac{i \cdot j}{4M \cdot F} \right) \right). \qquad (A4.5)$$

Expanding eqn A4.5 and using the relationships defined in A4.3 and A4.4 we get:

$$\delta = \prod_{l=1}^{L} \left(\begin{array}{l} \left(\sum_{i=0}^{2M} \binom{2M}{i} (F_{dad}^l)^i \cdot (1 - F_{dad}^l)^{2M-i} \cdot \frac{i}{2M} \cdot \right) \\ \left(\sum_{j=0}^{2F} \binom{2F}{j} (F_{dad}^l)^j \cdot (1 - F_{dad}^l)^{2F-j} \right) + \\ \sum_{j=0}^{2F} \binom{2F}{j} (F_{dad}^l)^j \cdot (1 - F_{dad}^l)^{2F-j} \cdot \frac{j}{2F} \cdot \\ \left(\sum_{i=0}^{2M} \binom{2M}{i} (F_{dad}^l)^i \cdot (1 - F_{dad}^l)^{2M-i} \right) - \\ \sum_{i=0}^{2M} \binom{2M}{i} (F_{dad}^l)^i \cdot (1 - F_{dad}^l)^{2M-i} \cdot \frac{i}{2M} \cdot \\ \left(\sum_{j=0}^{2F} \binom{2F}{j} (F_{dad}^l)^j \cdot (1 - F_{dad}^l)^{2F-j} \cdot \frac{j}{2F} \right) \end{array} \right)$$

$$= \prod_{l=1}^{L} \left(\frac{F_{dad}^l \cdot 2M}{2M} \cdot 1 + \frac{F_{dad}^l \cdot 2F}{2F} \cdot 1 - \frac{F_{dad}^l \cdot 2M}{2M} \cdot \frac{F_{dad}^l \cdot 2F}{2F} \right)$$

$$= \prod_{l=1}^{L} \left(2F_{dad}^l - (F_{dad}^l)^2 \right)$$

$$= NG_{dad}. \tag{A4.6}$$

Therefore, regardless of the actual number of mothers (M) or fathers (F) or the number of NGIs sampled, the expected proportion of the ($1 - Pat$) NGIs that are compatible with the putative father is simply NG_{dad}. Eqn A4.6 assumes that the M and F parents represent a random sample of the breeding population (with respect to the marker alleles) and that there is no mutation. Substituting eqns A4.2 and A4.6 into eqn A4.1 we get:

$$\overline{Pat} = \frac{Pat + (1 - Pat) \cdot NG_{dad} - N\hat{G}_{dad}}{1 - N\hat{G}_{dad}}. \tag{A4.7}$$

If $N\hat{G}_{dad} = NG_{dad}$ then $\overline{Pat} = Pat$. That is, if the estimated allele frequencies are accurate then the expected paternity is equal to the actual paternity and therefore, the paternity estimate is unbiased. Error in the allele frequency estimates can lead to a small bias (underestimate) in the parentage calculations. However, with only moderate sampling regimes, accurate allele frequencies are obtained and the bias is negligible. Consider the following example. Suppose that a paternity estimate was based on three loci, each with three equally common alleles, and the putative father was homozygous at each locus and had a paternity of 70%. If the allele frequencies were based on a sample of 20 or 100 random individuals, then nine out of 10 times the bias does not exceed 4.5% or 2.0%, respectively, and the expected bias is only 0.5% or 0.1%, respectively. From our bluegill example, we found that nine out of 10 times the bias did not exceed 0.7% and the expected bias is about 0.02%. If we had based the allele frequency estimates on only 20 individuals instead of 232, then nine out of 10 times the bias would not have exceeded 2.5% and the expected bias would be 0.3%. Finally, we found that as the allele frequencies in the putative parent's genotype decrease, the bias decreases (data not shown). Therefore, selecting loci with the greatest resolving power for specific parents will minimize the bias.