Molecular Ecology (2000) **9**, 529–539

Statistical confidence in parentage analysis with incomplete sampling: how many loci and offspring are needed?

B. D. NEFF,* J. REPKA† and M. R. GROSS*
*Department of Zoology, †Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 3G5

Abstract

We have recently presented models to estimate parentage in breeding systems with multiple mating and incomplete sampling of the candidate parents. Here we provide formulas to calculate the statistical confidence and the optimal trade-off between the number of loci and offspring. These calculations allow an understanding of the statistical significance of the parentage estimates as well as the appropriate sampling regime required to obtain a desired level of confidence. We show that the trade-off generally depends on the parentage of the putative parents. When parentage is low, sampling effort should concentrate on increasing the number of loci. Otherwise, there are similar benefits from increasing the number of loci or offspring. We demonstrate these methods using genetic data from a nest of the bluegill sunfish (*Lepomis macrochirus*).

Keywords: bluegill, genetic markers, maternity, parentage analysis, paternity, statistical confidence

Received 14 July 1999; revision received 26 October 1999; accepted 26 October 1999

Introduction

Parentage models are necessary to assess reproductive success, kinship and fitness of wild populations (Avise 1994; Jarne & Lagoda 1996; Petrie & Kempenaers 1998). Statistical confidence is an important parameter in parentage models because it reveals the accuracy of the estimates and thus the reliability of parentage inference (e.g. Pena & Chakraborty 1994; Evett & Weir 1998). Studies addressing statistical confidence in parentage estimates have focused on models that assume complete sampling of candidate parents and assign an offspring to the most-likely parent or parent pair (e.g. Chakraborty *et al*. 1974; Meagher 1986; Thompson 1986; Thompson & Meagher 1987; Evett & Weir 1998). Statistical confidence in the assignment of an offspring reflects the probability of correctly identifying the genetic parents, and generally increases as the number of loci increases and the number of candidate parents decreases (Chakraborty *et al*. 1988; Double *et al*. 1997; Estoup *et al*. 1998; Marshall *et al*. 1998). Statistical confidence estimators for parentage models that do not require the complete sampling of candidate parents are less available and are therefore necessary.

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

We have recently developed models for calculating the parentage of individuals in breeding systems with single-sex or two-sex multiple mating and when there is incomplete sampling of the candidate parents (Neff *et al*. 2000). Our models only require genetic data from the parent or parents in question, a sample of the next-generation individuals (NGIs) and an estimate of the breeding-population allele frequencies. They estimate the proportion of NGIs that are fathered or mothered by a putative parent. The models have been shown to provide unbiased estimates, accommodate loci with many alleles and be robust to violations of their assumptions. This makes the models particularly useful for providing parentage estimates when large sample sizes are analysed and when genetic data are available from only some of the candidate parents. We now provide an approach to calculate the statistical confidence associated with the parentage estimates.

Parentage analysis involving large numbers of NGIs must consider the number of NGIs to sample in addition to the number of loci. Increasing either will generally increase the confidence in the parentage estimates. This reflects a trade-off in the total number of genotypes that are analysed. Most laboratories would like to maximize their productivity by optimizing the trade-off between the number of NGIs and the number of loci. We therefore show how our statistical confidence calculations can be used to determine the optimal number of NGIs and loci for sampling.

We begin by developing formulas for calculating statistical confidence for the models of Neff *et al.* (2000). We show how these formulas can be applied to determine the optimal number of NGIs and loci to sample to achieve a desired level of confidence. We also show their optimal trade-off to minimize the number of genotypes. Finally, we demonstrate the application of the formulas using genetic data from a nest of the bluegill sunfish (*Lepomis macrochirus*). Bluegill sunfish have males that build nests and provide parental care and an alternative cuckolder life history that specializes at stealing fertilizations with the multiple females that may spawn in a nest.

Materials and methods

Statistical confidence

In Appendix I we derive the formulas for calculating statistical confidence in the models of Neff *et al.* (2000). Briefly, the variance in a parentage estimate is dependent on the frequency of the putative parent's genotype within the breeding population and the variance in the observed proportion of the NGIs that are genetically compatible with the putative parent or parents (i.e. ng_{dad}, ng_{mom}, or ng_{pair}; see Neff *et al.* 2000). This latter variance has two components. First, sampling error is introduced if not all of the NGIs are analysed. Second, there is variance from Mendelian inheritance, which follows the binomial distribution.

The confidence formulas require five parameters: (i) the number of NGIs sampled; (ii) the total number of NGIs (e.g. total size of a brood); (iii) the population frequency of the putative parent's or parents' alleles; (iv) the paternity, maternity, or parentage of the putative parent or parents; and (v) the effective number of breeders (other than the putative parent or parents) contributing genetically to the NGIs. The 'effective number' is the total number reduced by the variance in their success and is analogous to N_e in population genetics (e.g. Kimura 1983). Generally, several of these parameters will be unknown and must therefore be estimated. We compare estimation methods in the Discussion. The confidence formulas can also be used to calculate the 95% confidence interval (CI) (see Appendix I). The CI is useful because the variance may be asymmetrically distributed around the estimate.

In Appendix II we derive formulas for calculating the expected values of NG_{dad} or NG_{mom}, NG_{pair}, and NG_{pair}^{mepf} or NG_{pair}^{depf} . The ability of the models to make precise estimates is dependent on the values of these parameters, which are influenced by the number of loci and their polymorphism (the number of alleles).

How many loci and offspring?

The number of loci and the number of NGIs are typically the only two parameters of the five that can be manipulated in the confidence formulas (see the Discussion). Their product determines the total number of genotypes that are analysed (total genotypes = number of loci × number of NGIs). We use the Two-Sex Paternity model as an example of how to optimize the trade-off between these two parameters and minimize genotype number (the method is applicable to all of the parentage models). In the example, we assume that six males have contributed to fertilization of the brood. The putative father has a paternity of 80% and the remaining 20% are divided equally among the five additional fathers. We also assume that there are a total of five females, each contributing equally (20%). Thus, the brood includes genes from 11 parents that may have mated in all possible combinations of male and female pairs ($n = 30$ possible combinations). Using eqn A1.2 (Appendix I), the total variance in paternity is expressed as a function of both the putative father's genotypic frequency in the population (NG_{dad}; e.g. range = 0–0.5) and the number of NGIs analysed (e.g. 30 or 50). From these relationships we determine the combinations of the number of loci (as measured by NG_{dad}) and the number of NGIs that provide a desired variance or level of confidence. The optimal trade-off between these two numbers is then determined by the values that minimize the total number of genotypes (see the Results).

Biological example

The Two-Sex Paternity model was applied to genetic data to estimate the paternity of a parental male bluegill (Neff *et al.* 2000). Seven estimates of paternity were made, based on three microsatellite loci individually ($n = 3$), in pairs ($n = 3$) and all combined ($n = 1$). We then calculated the confidence associated with these estimates.

First, we assumed the paternity estimate that was based on all three loci to be the actual paternity of the putative father (83.6%; see Table 1). This estimate should be the most precise because it is based on the greatest amount of genetic information. Second, as the number of NGIs within the putative father's nest was much greater than in our sample, we used the binomial approximation to calculate the sampling error (see Appendix I). Third, as we did not know the effective number of breeders contributing genetically to the NGIs, we considered three possibilities: (i) only one genetic mother and one genetic cuckolder father; (ii) a minimum number of mothers and cuckolder fathers based on the genetic data; and (iii) the average population ratios of breeding females and cuckolder males to parental males. By assuming one genetic mother and cuckolder father we set the weakest level of

Table 1 The paternity results for the parental male bluegill

				Variance (%)		
Loci	NG_{dad}	ng_{dad} (%)	*Pat* (%)	1 ♀, 1 P♂, 1 C♂	2 ♀, 1 P♂, 1 C♂	4 ♀, 1 P♂, 6 C♂
Single						
Lma102	0.551	89.1 (41/46)	75.8	2.2	1.8	1.0
Lma120	0.563	91.3 (42/46)	80.1	2.3	1.8	1.0
Lma87	0.658	93.5 (43/46)	80.9	3.1	2.5	1.4
Paired						
Lma102, Lma120	0.310	89.1 (41/46)	84.2	0.91	0.77	0.53
Lma102, Lma87	0.362	87.0 (40/46)	79.6	1.1	0.91	0.60
Lma120, Lma87	0.370	87.0 (40/46)	79.3	1.1	0.93	0.61
All						
Lma102, Lma120	0.204	87.0 (40/46)	83.6	0.61	0.54	0.42
Lma87				(31–93)	(39–93)	(65–93)
Range	0.204–0.658	87.0–93.5	75.8–84.2	0.61–3.1	0.54–2.5	0.42–1.4

NG_{dad} and ng_{dad} are defined in Neff *et al*. (2000). *Pat* is the paternity estimate for the parental male. Three estimates of the variance associated with *Pat* are provided. The first assumes that there is one mother and two males (one parental and one cuckolder); the second assumes that there are two effective mothers and two males; the third assumes that there are four effective mothers and seven males (one parental and six effective cuckolders). The 95% confidence interval is included for the estimates based on all loci. The range in values is also indicated.

confidence in the paternity estimate because confidence decreases with fewer breeders (see the Results). The genotypes of the 46 NGIs at the three loci require a minimum of two maternal genotypes (see Table 5 from Neff *et al*. 2000; e.g. Mother 1: 98/102; 217/227; 128/152; and Mother 2: 98/102; 211/231; 128/152 at Lma102, 120 and 87, respectively) and one cuckolder genotype in addition to the parental male (e.g. Cuckolder 1: 98/102; 231/245; 118/128 at Lma102, 120 and 87, respectively). The effective number of parents may be more than the minimum numbers because the parents may have similar or identical genotypes and may therefore be undetectable. Furthermore, the sample of NGIs may not contain offspring from all the genetic parents (the absolute maximum number of parents would be a different parent for each NGI). In both these cases, the minimum number of parents would be a conservative estimate. Alternatively, the effective number of parents may be less than the minimum numbers when reproductive success among the parents is highly skewed (the absolute minimum would approach one genetic mother and one genetic cuckolder father). In this instance, the minimum number of parents would overestimate the effective number. The average ratios of breeders allowed us to calculate the expected level of confidence based on biological data for the population at large, and should provide the most accurate estimate of confidence because it reflects a probable number of females and cuckolder males to spawn in the putative father's nest. However, if their reproductive success is highly skewed then it can overestimate the effective number of breeders.

We also calculated the 95% CI in the paternity estimate based on the three possibilities for the effective number of breeders. We used eqn A1.29 (Appendix I) to generate a distribution of the probability of observing ng_{dad} (the observed proportion of offspring that are compatible with the putative father) given the putative father's genotype, over the range of possible *Pat* values (range = 0–1). The distribution was normalized and the 95% CI was numerically calculated from the areas under the curve representing 2.5 and 97.5%.

Results

Statistical confidence

Four factors influence the variance, and hence confidence, in the estimates made by the parentage models of Neff *et al*. (2000). We present detailed results for the Two-Sex Paternity model (the relationships are similar for all the models). First, the frequency of the putative parent's genotype within the breeding population, as measured by the *NG* parameters, directly influences the variance (Fig. 1a,b,c). *NG* increases as the number of loci and their degree of polymorphism decreases. As *NG* increases, the variance in the estimate also increases. The effect of a small change in *NG* is greatest when it is large, when few NGIs are analysed, and when the putative parent's fertilization success is low. At an *NG* value of 0, the putative parent's offspring could be unambiguously identified and only sampling error introduces variance into the estimate. Second, as the number of NGIs in a sample increases, the variance decreases (Fig. 1a). This is largely a result of reduced sampling error. The effect of a small change in the number of NGIs is greatest when only a few are sampled and when *NG* is large. Third, as the proportion of NGIs that are produced by the putative parent (e.g. paternity) increases, the variance decreases (Fig. 1b). The effect of a small change in paternity is greatest when paternity is low and when *NG* is large. Paternity also influences sampling error, but has a significant effect only at very low values of *NG*. Fourth, as the number of individuals that genetically contribute to the NGIs increases, the variance decreases (Fig. 1c). The effect of a small change in number of effective breeders is greatest when there are few and when *NG* is large. When *NG* is zero, the number of effective breeders has no effect on the variance.

The expected value of *NG* decreases as the number of alleles or the number of loci increases (Fig. 2; Appendix II). With only a moderate number of polymorphic loci, low values of *NG* are obtained. For example, the corresponding values of NG_{dad} or NG_{mom}, NG_{pair} and NG_{pair}^{mepf} or NG_{pair}^{depf} for one locus with five equally common alleles are 0.5840, 0.2282 and 0.4048, respectively. With five such loci these values decrease to 0.0679, 0.0006 and 0.0109, respectively.

How many loci and offspring?

Increasing the number of NGIs or decreasing *NG* both decrease the variance. An example of the trade-off in the number of loci (as measured by *NG*) and the number of offspring (NGIs) in determining the variance in a paternity estimate is shown in Fig. 3. A desired variance of 0.009 could be achieved with 30 NGIs and an NG_{dad} of ≈ 0.27 or with 50 NGIs and an NG_{dad} of ≈ 0.44. Suppose that we could obtain an NG_{dad} value of 0.44 and 0.27 with one locus and two loci, respectively. As in the first instance, examining 30 NGIs with two loci would require 60 genotypes, while, as in the second instance, examining 50 NGIs with one locus would only require 50 genotypes. Both approaches provide an estimate with the same variance and therefore confidence. However, the second approach would require fewer genotypes and would thus be more efficient.

As a second example, suppose that there are genotypes from 30 NGIs with sufficient loci to obtain an NG_{dad} value of 0.27 and that it is desirable to decrease the variance from 0.009 to 0.006. This could be accomplished by either

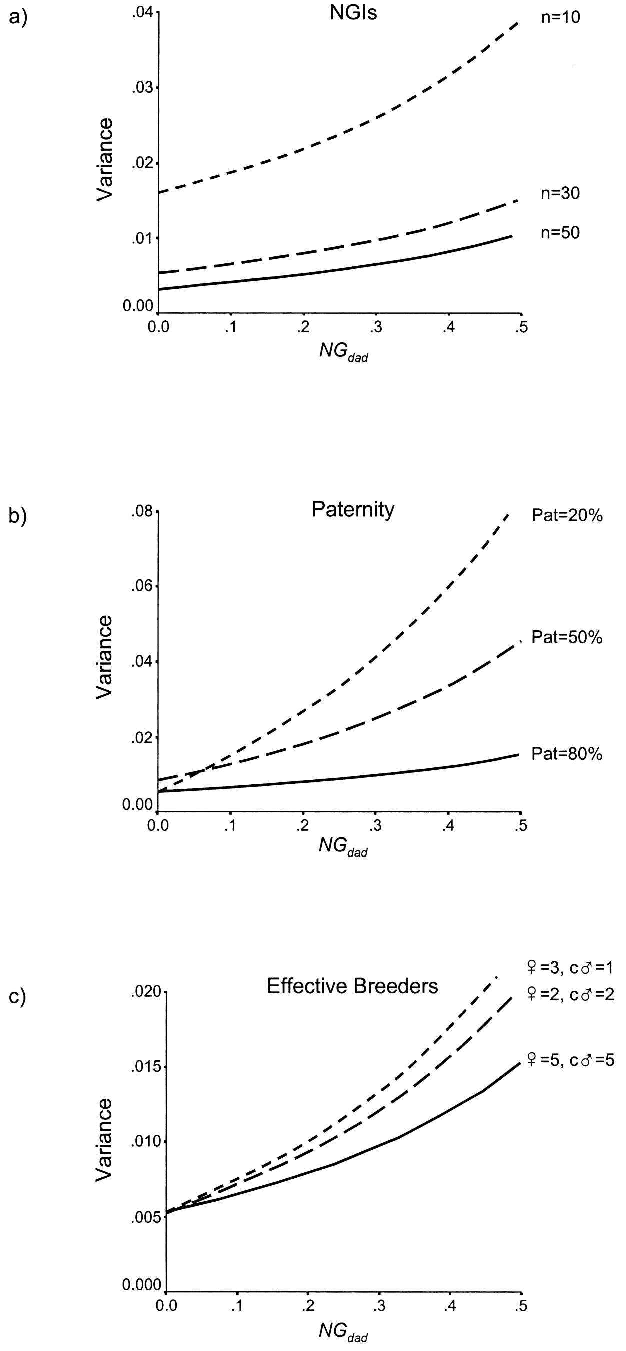

Fig. 1 The variance in an estimate of paternity is effected by the value of NG_{dad}, the number of next-generation individuals (NGIs), paternity and the effective number of breeders, as shown. The relationships are derived from the Two-Sex Paternity model using eqn A1.2, but they are also applicable to all the models. (a) The variance in the estimate decreases as the number of NGIs increases or as the value of NG_{dad} decreases. The expected variance is shown for three sample sizes of NGIs (solid line = 50 NGIs; long-dashed line = 30; short-dashed line = 10). At an NG_{dad} value of zero, the entire variance a result of sampling error. In this example, the paternity of the putative father was 80% and five cuckolder males and five females contributed genetically to the brood. (b) The variance in the estimate decreases as paternity increases (except at very low values of NG_{dad} where sampling error is the major source of variance). The expected variance is shown for three values of paternity (Pat) (solid line = 80%; long-dashed line = 50%; short-dashed line = 20%). In this example, 30 NGIs were analysed and five cuckolder males and five females contributed to the brood. (c) The variance in the estimate decreases as the effective number of breeders increases. The solid line represents 10 effective breeders, five cuckolder males and five females, in addition to the putative father. The dashed lines represent four effective breeders in addition to the putative father (short-dashed line: one cuckolder male and three females; long-dashed line: two cuckolder males and two females). In this example, 30 NGIs were analysed and the paternity of the putative father was 80%.

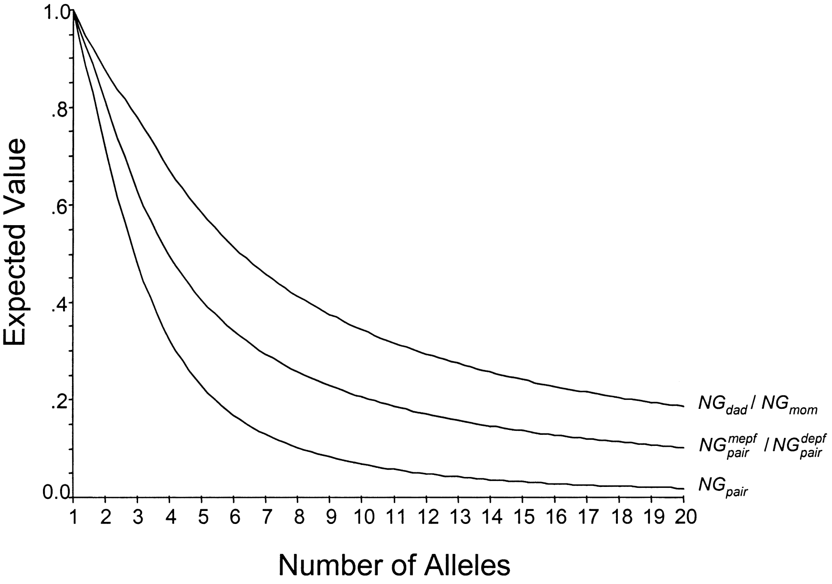

Fig. 2 The expected value of *NG* decreases as the number of alleles increases, as shown for NG_{dad} or NG_{mom}, NG_{pair}^{mepf} or NG_{pair}^{depf} and NG_{pair} for loci with one to 20 equally common alleles.

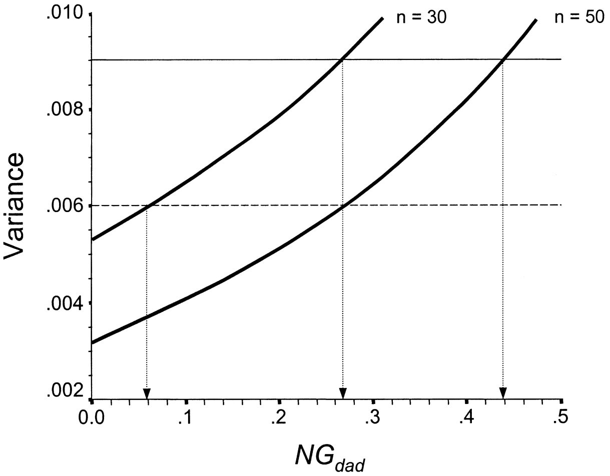

Fig. 3 An example of how the trade-off between the number of next-generation individuals (NGIs) and the number of loci can be assessed. The two heavy solid lines represent the relationship between the total variance in the paternity estimate and the value of NG_{dad} for 30 and 50 NGIs. As an example (also see the text), a variance of 0.006 could be obtained by analysing 30 NGIs with sufficient loci to obtain a NG_{dad} value of ≈ 0.06 or by analysing 50 NGIs with fewer loci having a NG_{dad} value of only ≈ 0.27. The optimal approach would probably require the generation of the fewest genotypes. The lines were derived using eqn A1.2 assuming five females and five cuckolder males and a paternity for the putative father of 80%.

examining an additional 20 NGIs (50 total) with the same loci used to genotype the initial 30 or obtaining genotypes from the initial 30 NGIs with additional loci to reduce the value of NG_{dad} to ≈ 0.06 (see Fig. 3). An optimal approach would require the least additional effort and cost. Suppose that two loci were used to obtain the genotypes from the initial 30 NGIs and that it would require another two loci to obtain a NG_{dad} value of 0.06. The first option, of running 20 additional NGIs, would require the generation of only 40 additional genotypes (20 NGIs × two loci). The second option, of running the initial 30 NGIs with the additional two loci, would require the generation of 60 additional genotypes (30 NGIs × two loci). Again, both approaches provide the same level of confidence, but the first would require fewer additional genotypes and would thus be more efficient.

Biological example

Table 1 presents the variances in the seven paternity estimates for the parental male bluegill. The variance decreases with decreasing value of NG_{dad} (i.e. increasing resolving power of the loci) and as the effective number of breeders increases. As expected, the paternity estimate based on all three loci was the most precise. Furthermore, the estimate under the assumption that there is only one effective mother and one effective cuckolder father had the greatest variance and therefore was the most conservative. The estimate under the assumption of two effective mothers and one effective cuckolder was the next most conservative, and the estimate under the assumption of four effective mothers and six effective cuckolder fathers was the least conservative. However, the latter estimate may be the most accurate because in Lake Opinicon the population breeding ratios are approximately four females and six cuckolder males to each parental male (Gross 1982, 1991). Therefore, based on all three loci, we conclude that the parental male bluegill has a probable paternity of 83.6 ± 7% (SD; variance = 0.42%; 95% CI: 65–92%).

Discussion

We have presented formulas to calculate the statistical confidence associated with the parentage estimates made using the models of Neff *et al.* (2000). These models estimate the proportion of offspring fathered or mothered by a putative parent and are particularly useful for analysing large sample sizes. When analysing parentage of a large number of individuals, sampling regimes for both offspring and loci are necessary. Therefore, we have also presented methods that use the confidence formulas to determine the optimum number of offspring and loci to achieve a desired level of statistical confidence.

Studies addressing statistical confidence in parentage estimates have focused on models that assume complete sampling of candidate parents. These models attempt to exclude all but the genetic parents or assign an offspring to the most probable nonexcluded pair (e.g. Chakraborty *et al.* 1974; Meagher 1986; Thompson 1986; Thompson & Meagher 1987; Evett & Weir 1998). Confidence statistics for these models calculate the probability of identifying the true parents and emphasize the number of genetic loci needed (Chakraborty *et al.* 1988; Double *et al.* 1997; Estoup *et al.* 1998; Marshall *et al.* 1998). By contrast, the confidence statistics for our models calculate the variance associated with the estimated proportion of offspring fathered or mothered by a putative parent, and emphasize the optimal numbers of both loci and offspring.

The confidence of the parentage estimates is influenced by both the number of NGIs analysed and the number of genetic loci used. These factors represent a trade-off in the total number of genotypes analysed. As such, most laboratories would like to maximize their productivity by optimizing this trade-off. We provide the first methods to determine the optimal number of NGIs and loci needed to obtain parentage estimates with a desired level of confidence. Generally, when the putative parent has low reproductive success (< 25%) the emphasis is on implementing additional loci, especially when at least 30 NGIs are analysed. However, at higher success (> 80%) the emphasis may be on increasing either the number of NGIs or the number of loci. The confidence formulas enable researchers to determine the minimum number of NGIs and loci needed to obtain parentage estimates with a desired level of confidence. As such, these methods should greatly increase the efficiency of parentage analysis.

The confidence of the parentage estimates also depends on the putative parent's reproductive success and the effective number of breeders that contribute to a sample of NGIs. Generally, neither of these parameters can be manipulated and therefore nothing can be done with them to increase confidence. However, all else equal, estimates for putative parents with higher reproductive success or from samples of NGIs that are produced by a greater number of effective breeders will have greater confidence. These results can be used to provide more precise parentage estimates. For example, suppose that for a sample of NGIs there was a single genetic mother and two potential genetic fathers (e.g. a parental male and a cuckolder male), and that we could catch only one of these males. As there are only two males, their paternities (expressed as a proportion) must total to one. Based on biological data (e.g. behavioural data), suppose that we anticipate that the parental male will have higher paternity than the cuckolder male. If we can collect only one male, we should collect the parental male and estimate his paternity and use it to calculate the cuckolder's paternity (i.e. one less the paternity of the parental male) because this will maximize the precision of the estimates.

The confidence statistics developed here require five parameters, of which several may be unknown. First, the total number of NGIs is sometimes too large to count accurately (e.g. the broods of many fish can contain tens of thousands of NGIs). However, the total number of NGIs is not required if it is much larger than the number sampled because the sampling error can be calculated with accuracy (Zar 1999). By contrast, if the total number of NGIs is not much larger than the sample (but is unknown) then the binomial approximation will overestimate the true sampling error and will thereby provide a conservative estimate of the true confidence.

Second, the frequency of a putative parent's alleles will often be estimated from a sample of the breeding population. The sampling error in the allele frequencies can introduce a small bias into the parentage estimates, but does not introduce variance into the confidence estimates (see Neff *et al.* 2000 and Appendix I). Therefore, although sampling error in the allele frequency estimates should be minimized to mitigate a potential bias, it does not directly influence the calculation of confidence.

Third, we have suggested that the calculated paternity, maternity or parentage can be used as an estimate of its actual value. The calculations from the models provide accurate estimates (see Neff *et al.* 2000) and if the putative parent or parent pair has a rare multilocus genotype then the estimate is also precise. In such cases, the true value will not vary much from this estimate. Furthermore, small differences in the value used for the actual paternity, maternity or parentage do not substantially influence the predicted variance when NG is small. As an example, consider the bluegill sunfish. Based on only three loci (NG_{dad} = 0.204) we estimated the paternity of the parental male to be 83.6 ± 7% (95% CI: 65–92%). If we had assumed that the actual paternity of the parental male was 65% (lower 65% CI) then the associated standard deviation in the estimate would be only marginally higher, at 9%. Therefore, using the estimate as the actual value should not influence the predicted variance significantly, especially when *NG* is small. Furthermore, the CI calculation does not require the actual reproductive success of the putative parent or parents.

Finally, we described three approaches to estimate the effective number of breeders. The simplest and most conservative approach is to assume that there is only one of each. In this case, the calculation provides a minimum level of confidence (i.e. a maximal estimate of the variance), and can significantly underestimate the true level of confidence. For example, consider the bluegill sunfish. Using the population breeding ratios of six cuckolder males and four females per parental male, we determined that the 95% CI was 65–92%. However, if we assume that there was only one cuckolder male and one female then the CI is considerably larger, at 31–93%. Therefore, researchers may wish to consider both the most conservative estimate of the effective number of breeders and potentially better estimates, such as the population breeding ratios.

It is possible that the robustness of the parentage estimates could be assessed in a more *ad hoc* manner using bootstrap approaches in subsequent statistical analyses involving the estimates. However, this approach would not allow the calculation of the optimal trade-offs in the number of loci and offspring. Furthermore, the confidence statistics presented here allow partitioning of the parentage data according to precision, which can in turn be used to increase statistical power in subsequent analyses.

We have previously made seven estimates of the paternity of the parental male bluegill to the brood within his nest. These estimates were based on three microsatellite loci treated individually, in pairs and collectively. Although we suspected that the estimate based on all three loci was the most precise, we did not known how precise it was. We now know that of the seven estimates it is, in fact, the most precise and that the variance is two to three times lower than for the three loci used individually. We therefore conclude that the paternity of the parental male bluegill is 83.6 ± 7% (95% CI: 65–92%).

Acknowledgements

Our work is supported by the Natural Science and Engineering Research Council of Canada.

References

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

Chakraborty R, Shaw M, Schull WJ (1974) Exclusion of paternity: the current state of the art. *American Journal of Human Genetics*, **26**, 477–488.

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

Double MC, Cockburn A, Barry SC, Smouse PE (1997) Exclusion probabilities for single-locus paternity analysis when related males compete for matings. *Molecular Ecology*, **6**, 1155–1166.

Estoup A, Gharbi K, SanCristobal M, Chevalet C, Haffray P, Guyomard R (1998) Parentage assignment using microsatellites in turbot (*Scophthalmus maximus*) and rainbow trout (*Oncorhynchus mykiss*) hatchery populations. *Canadian Journal of Fisheries and Aquatic Sciences*, **55**, 715–725.

Evett IW, Weir BS (1998) *Interpreting DNA Evidence*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA.

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: Biological Sciences*, **332**, 59–66.

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

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

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.

Neff BD, Repka J, Gross MR (2000) Parentage analysis with incomplete sampling of candidate parents and offspring. *Molecular Ecology*, **9**, 515–528.

Pena SDJ, Chakraborty R (1994) Paternity testing in the DNA era. *Trends in Genetics*, **10**, 204–209.

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

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

Thompson EA (1986) Likelihood inference of paternity. *American Journal of Human Genetics*, **39**, 285–287.

Thompson EA, Meagher TR (1987) Parental and sib likelihoods in genealogy reconstruction. *Biometrics*, **43**, 585–600.

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

Appendix I

Derivation of the variance associated with estimates of paternity (*Pat*), maternity (*Mat*) and parentage (*Par*) in Neff *et al*. (2000).

The derivation for the confidence in the Two-Sex Paternity model is presented in detail. The remaining models are presented more concisely. All variables not defined here are defined in Neff *et al*. (2000).

Two-Sex Paternity or Maternity

Here we derive the confidence for the Two-Sex Paternity model; the derivation for the Two-Sex Maternity model is analogous. From Neff *et al*. (2000):

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

The variance in the estimate *Pat* can be expressed as:

$$\text{var}\,(Pat) = \text{var}\left(\frac{ng_{dad} - NG_{dad}}{1 - NG_{dad}}\right). \tag{A1.1}$$

NG_{dad} is a constant for a given putative father's genotype (see Neff *et al*. 2000) and therefore:

$$\text{var}\,(Pat) = \frac{\text{var}\,(ng_{dad})}{(1 - NG_{dad})^2}. \tag{A1.2}$$

If we assume that *n* NGIs have been sampled and analysed from a total of *N* NGIs, then the variance in the proportion of NGIs that are compatible with the putative father can be calculated from:

$$\text{var}\,(ng_{dad}) = \sum_{k=0}^{n}\left(\text{Pr}(k)\cdot\left(\overline{ng}_{dad} - \frac{k}{n}\right)^2\right). \tag{A1.3}$$

Here, Pr(*k*) is the probability that *k* of the *n* NGIs are compatible with the putative father and $\overline{ng}_{dad}$ is the expected proportion of the *n* NGIs that are compatible with the putative father, and is calculated from:

$$\overline{ng}_{dad} = \sum_{k=0}^{n}\left(\text{Pr}(k)\cdot\frac{k}{n}\right). \tag{A1.4}$$

The probability Pr(*k*) has two components, and is calculated from:

$$\text{Pr}(k) = \sum_{n_1=0}^{n}(\text{Pr}(\mathbf{n})\cdot Pr(k\,|\,\mathbf{n})). \tag{A1.5}$$

The first component is a sampling probability that is dependent on the vector $\mathbf{n} = (n_1, n_2 = n - n_1)$, where n_1 and n_2 are the numbers of the *n* NGIs that are produced by the putative father and other fathers, respectively. The sampling probability depends on the number of NGIs sampled (*n*), the total number of NGIs for which the paternity estimate is being made (*N*) and the paternity of the putative father (*Pat*), and can be calculated from:

$$\text{Pr}(\mathbf{n}) = \binom{n}{n_1}\cdot\frac{N_1!\cdot N_2!\cdot(N-n)!}{(N_1-n_1)!\cdot(N_2-n_2)!\cdot N!}; \tag{A1.6}$$

where

$$N_1 = Pat\cdot N; \tag{A1.7}$$

$$N_2 = (1 - Pat)\cdot N = N - N_1. \tag{A1.8}$$

Note that N_1 and N_2 will be integers because *Pat* is defined as the proportion of the *N* NGIs that belong to the putative father and therefore can only take on the values of $0/N$, $1/N$, … , or N/N. As an example, if 50 offspring were analysed from a brood of 1000 and the putative father had a paternity of 80% then $n = 50$, $N = 1000$, $N_1 = 800$ and $N_2 = 200$. If *N* is much larger than *n* (e.g. if *n* is not > 5% of *N*) then eqn A1.6 can be approximated by the binomial theorem (Zar 1999):

$$\text{Pr}(\mathbf{n}) \approx \binom{n}{n_1}\cdot Pat^{n_1}\cdot(1-Pat)^{n_2}. \tag{A1.9}$$

Note that this equation is independent of the total number of NGIs (i.e. *N*).

The second component represents the probability that *k* of the *n* offspring are compatible with the putative father given that n_1 of them are actually produced by him. It is dependent on the number of effective mothers (*M*), the number of effective fathers excluding the putative father (*F*), and the frequency of the alleles in the genotype of the putative father. It is calculated from:

$$\text{Pr}(k\,|\,\mathbf{n}) = \sum_{\mathbf{i},\mathbf{j}}\left(\begin{matrix}\displaystyle\prod_{l=1}^{L}\left(\binom{2F}{i_l}\binom{2M}{j_l}(F_{dad}^{l})^{i_l+j_l}\cdot(1-F_{dad}^{l})^{2F+2M-i_l-j_l}\right)\cdot\\ \displaystyle\binom{n_2}{k-n_1}(C_{dad})^{k-n_1}\cdot(1-C_{dad})^{n-k}\end{matrix}\right); \tag{A1.10}$$

where the probability that a compatible NGI is produced from a mating between one of the *M* mothers and *F* fathers given **i** and **j** is:

$$C_{dad} = \prod_{l=1}^{L}\left(\frac{i_l}{2F} + \frac{j_l}{2M} - \frac{i_l\cdot j_l}{4F\cdot M}\right). \tag{A1.11}$$

The first line of eqn A1.10 represents the probability that i_l of the 2*F* paternal alleles and j_l of the 2*M* maternal alleles are shared with the putative father at each of the *L* loci. The second line represents the probability that k–n_1 of the n_2 NGIs are compatible (i.e. inherit at least one allele that is shared) with the putative father given **i** of the paternal and **j** of the maternal alleles are shared with him. The summation is over all combinations of i_ls and j_ls satisfying $i_l \le 2F$ and $j_l \le 2M$ for every *l*. Eqn A1.10 assumes that the number of effective fathers (*F*) and mothers (M) are whole numbers. We have also developed a formula that incorporates the variance in reproductive success among genetic parents and therefore allows a fractional number of effective breeders. However, it is not presented here as it is unlikely that both the number of genetic parents and the variance in their reproductive success will be known and therefore its usefulness is limited. It can be shown that eqn A1.10 follows from this more complex formula with the assumption that each of the F fathers and each of the M mothers have equal reproductive success (see the Discussion for methods to estimate F and M).

Two-Sex Parentage

The derivation of the Two-Sex Parentage model follows an analogous approach to the previous model. Here we derive the variance associated with the parentage estimate assuming that $Par = Pat = Mat$ (i.e. assuming that the putative parents have their entire success with each other). The derivation for other cases is considerably more elaborate and is not presented here. Given $Par = Pat = Mat$, the variance in Par can be calculated from:

$$\mathrm{var}\,(Par) = \frac{\mathrm{var}\,(ng_{pair})}{(1-NG_{pair})^2}; \qquad \text{(A1.12)}$$

The variance in ng_{pair} can be calculated from:

$$\mathrm{var}\,(ng_{pair}) = \sum_{k=0}^{n}\left(\Pr(k)\cdot\left(\overline{ng}_{pair}-\frac{k}{n}\right)^2\right). \qquad \text{(A1.13)}$$

The expected value of ng_{pair} can be calculated from:

$$\overline{ng}_{pair} = \sum_{k=0}^{n}\left(\Pr(k)\cdot\frac{k}{n}\right); \qquad \text{(A1.14)}$$

The probability Pr(k) is defined above (eqn A1.5). For this model the vector $\mathbf{n}$ contains two elements (n_1, $n_2 = 1-n_1$), where n_1 and n_2 are the numbers of the n NGIs that are produced by the putative parents and other parents, respectively. The first component of eqn A1.5, the probability of observing the allocation $\mathbf{n}$, can be calculated from:

$$\Pr(\mathbf{n}) = \binom{n}{n_1}\cdot\frac{N_1!\cdot N_2!\cdot(N-n)!}{(N_1-n_1)!\cdot(N_2-n_2)!\cdot N!}; \qquad \text{(A1.15)}$$

where

$$N_1 = Par\cdot N; \qquad \text{(A1.16)}$$

$$N_2 = (1-Par)\cdot N = N - N_1. \qquad \text{(A1.17)}$$

or if $N \gg n$ then:

$$\Pr(\mathbf{n}) \approx \binom{n}{n_1}\cdot Par^{n_1}\cdot(1-Par)^{n_2}. \qquad \text{(A1.18)}$$

The second component from eqn A1.5 can be calculated from:

$$\Pr(k\,|\,\mathbf{n}) = \sum_{\mathbf{x},\mathbf{y}}(\Pr(k\,|\,\mathbf{n},\mathbf{x},\mathbf{y})\cdot\Pr(\mathbf{x})\cdot\Pr(\mathbf{y})); \qquad \text{(A1.19)}$$

where

$$\Pr(k\,|\,\mathbf{n},\mathbf{x},\mathbf{y}) = \binom{n_2}{k-n_1}(C_{pair})^{k-n_1}\cdot(1-C_{pair})^{n-k}; \qquad \text{(A1.20)}$$

$$C_{pair} = \prod_{l=1}^{L}\left(\frac{x_{l1}}{2F}\cdot\left(\frac{y_{l2}+y_{l3}}{2M}\right)+\frac{x_{l2}}{2F}\cdot\left(\frac{y_{l1}+y_{l3}}{2M}\right)+\frac{x_{l3}}{2F}\cdot\left(\frac{y_{l1}+y_{l2}+y_{l3}}{2M}\right)\right); \qquad \text{(A1.21)}$$

$$\Pr(\mathbf{x}) = \prod_{l=1}^{L}\left(\frac{(2F)!}{x_{l1}!\cdot x_{l2}!\cdot x_{l3}!\cdot(2F-x_{l1}-x_{l2}-x_{l3})!}\cdot\left((Fu^{l}_{dad})^{x_{l1}}\cdot(Fu^{l}_{mom})^{x_{l2}}\cdot(Fs^{l}_{pair})^{x_{l3}}\cdot(1-Fu^{l}_{dad}-Fu^{l}_{mom}-Fs^{l}_{pair})^{2F-x_{l1}-x_{l2}-x_{l3}}\right)\right); \qquad \text{(A1.22)}$$

$$\Pr(\mathbf{y}) = \prod_{l=1}^{L}\left(\frac{(2M)!}{y_{l1}!\cdot y_{l2}!\cdot y_{l3}!\cdot(2M-y_{l1}-y_{l2}-y_{l3})!}\cdot\left((Fu^{l}_{dad})^{y_{l1}}\cdot(Fu^{l}_{mom})^{y_{l2}}\cdot(Fs^{l}_{pair})^{y_{l3}}\cdot(1-Fu^{l}_{dad}-Fu^{l}_{mom}-Fs^{l}_{pair})^{2M-y_{l1}-y_{l2}-y_{l3}}\right)\right); \qquad \text{(A1.23)}$$

and Fu^{l}_{dad} is the sum of the frequency of the putative father's unique alleles that are not shared with the putative mother at locus l; Fu^{l}_{mom} is analogous to Fu^{l}_{dad}.

Fs^{l}_{pair} is the sum of the frequency of the unique alleles shared by the putative father and putative mother at locus l.

In eqn A1.19 we must calculate the probability that k of the n NGIs are compatible with the putative parents. As only n_1 are produced by the putative parents, k–n_1 are compatible by chance and are produced by matings between females and males other than the putative parents.

The matrix $\mathbf{x}$ contains L (number of loci) rows and three columns. For a given row (locus l), the elements in the three columns (x_{l1}, x_{l2}, x_{l3}) are indices representing the number of the 2F effective paternal alleles that are equivalent to at least one of the putative father's but none of the putative mother's, at least one of the putative mother's but none of the putative father's, and at least one of the putative father's and at least one of the putative mother's, respectively. The matrix $\mathbf{y}$ is defined analogously for the 2M effective maternal alleles. The probability of a particular $\mathbf{x}$ or $\mathbf{y}$ distribution is calculated based on the binomial theorem and the frequency of the putative mothers' and fathers' shared and unshared alleles (see eqns A1.22 and A1.23). The summation in eqn A1.19 is over all combinations of x_{li}s and y_{li}s satisfying $\sum_i x_{li} \le 2F$ and $\sum_i y_{li} \le 2M$ for every l.

Single-Sex Paternity or Maternity

Here we derive the confidence for the Single-Sex Paternity model; the derivation for the Single-Sex Maternity model is analogous. The variance in the paternity estimate can be calculated from:

$$\mathrm{var}\,(Pat) = \frac{\mathrm{var}\,(ng_{pair})}{(1-NG_{pair}^{mepf})^2}; \tag{A1.24}$$

where

$$\mathrm{var}\,(ng_{pair}) = \sum_{k=0}^{n}\left(\Pr(k)\cdot\left(\overline{ng}_{pair}-\frac{k}{n}\right)^2\right). \tag{A1.25}$$

The expected value of ng_{pair} and $\Pr(k)$ are defined above (eqns A1.14 and A1.5, respectively). Here, $\Pr(k\,|\,\mathbf{n})$ is calculated from:

$$\Pr(k\,|\,\mathbf{n}) = \sum_{\mathbf{i},\mathbf{j}}\left(\begin{array}{l}\prod_{l=1}^{L}\left(\begin{array}{l}\left(\frac{(2F)!}{i_l!\cdot j_l!\cdot(2F-i_l-j_l)!}\right)^{i_l}\cdot(F_{dad}^{l})^{i_l}\cdot\\ (P_{lu})^{j_l}\cdot(1-F_{dad}^{l}-P_{lu})^{2F-i_l-j_l}\end{array}\right)\cdot\\ \binom{n_2}{k-n_1}(C_{pair})^{k-n_1}\cdot(1-C_{pair})^{n-k}\end{array}\right); \tag{A1.26}$$

where the probability that a compatible NGI is produced from a mating between the genetic mother and one of the F fathers given **i** and **j** is:

$$C_{pair} = \prod_{l=1}^{L}\left(\frac{i_l+\frac{1}{2}\cdot j_l}{2F}\right). \tag{A1.27}$$

The first line in eqn A1.26 represents the probability that i_l of the $2F$ paternal alleles are shared with the putative father and j_l of the $2F$ paternal alleles are equivalent to the genetic mother's unshared allele (with frequency P_{lu}) when she is heterozygous and shares exactly one allele with the putative father at locus l (see Neff *et al.* 2000). The second line represents the probability that k–n_1 of the n_2 NGIs are compatible with the putative parents given **i** and **j**. The summation is over all combinations of i_ls and j_ls satisfying $(i_l + j_l) \le 2F$ for every l.

Confidence Intervals

The formulas derived above can be used to calculate a confidence interval (CI). As an example, consider the Two-Sex Paternity model (the other models follow the same format). To calculate a CI we need to generate the distribution for the probability that a putative father has a paternity *Pat* given the observed ng_{dad} over the range of possible paternity values ($Pat = 0$–1). From Bayes' rule this can be calculated from:

$$\Pr(Pat\,|\,ng_{dad}) = \frac{\Pr(ng_{dad}\,|\,Pat)}{\Pr(ng_{dad})}\cdot\Pr(Pat). \tag{A1.28}$$

Recall that eqn A1.5 represents the probability that k of n offspring are compatible with the putative father given his paternity and therefore provides a formula to calculate $\Pr(ng_{dad}\,|\,Pat)$ (where $k = ng_{dad} \times n$). The probability $\Pr(ng_{dad})$ is a constant given the putative father's multilocus genotype and the effective number of parents that contribute to the NGIs, and therefore becomes part of the normalization constant (C) when the probability distribution is normalized such that the area under the curve (range $Pat = 0$–1) equals one. The *a priori* probability of a given paternity ($\Pr(Pat)$), however, is generally unknown. A conservative assumption is to assume that this probability follows a uniform distribution and is therefore a constant for any given *Pat* (Pena & Chakraborty 1994; Smouse & Meagher 1994; B. D. Neff *et al.* unpublished). In this case, $\Pr(Pat)$ also becomes part of the normalization constant and therefore we have:

$$\Pr(Pat\,|\,ng_{dad}) \approx C\cdot\Pr(ng_{dad}\,|\,Pat). \tag{A1.29}$$

As an example, the 95% CI is calculated from the distribution generated from eqn A1.29 by determining the values of *Pat* corresponding to areas of 2.5% and 97.5%.

The formulas derived in this appendix can be computationally intensive depending in the numbers of loci and effective breeders. However, these formulas can be easily evaluated using Monte Carlo simulations. In future work the authors will make software available.

Appendix II

Derivations of the expected values of NG_{dad} (NG_{mom}), NG_{pair}, and NG_{pair}^{mepf} (NG_{pair}^{depf}).

For a given locus and population, the expected value of NG_{dad} (NG_{mom} is analogous) is the sum of its values for each possible genotype weighted by the frequency of the genotype:

$$\overline{NG}_{dad} = \sum_{i=1}^{A}\sum_{j=1}^{A}(P_{li}P_{lj}\cdot NG_{ij}); \tag{A2.1}$$

where $\overline{NG}_{dad}$ is the expected value of NG_{dad} for a given locus; A is the number of alleles at the locus; and NG_{ij} is the value of NG_{dad} given the genotype consisting of alleles i and j, and is calculated as in Neff *et al.* (2000).

The expected value for multiple loci (when considered simultaneously) is the product of the expected NG_{dad} values for each locus. The expected value of NG_{pair} for a given locus is:

$$\overline{NG}_{pair} = \sum_{i=1}^{A}\sum_{j=1}^{A}\sum_{k=1}^{A}\sum_{m=1}^{A}(P_{li}\cdot P_{lj}\cdot P_{lk}\cdot P_{lm}\cdot NG_{ijkm}); \tag{A2.2}$$

where $\overline{NG}_{pair}$ is the expected value of NG_{pair} for a given locus; and NG_{ijkm} is the value of NG_{pair} given the father's genotype consisting of alleles i and j and the mother's genotype consisting of alleles k and m, and is calculated as in Neff *et al.* (2000).

The expected value for multiple loci (when considered simultaneously) is the product of the expected NG_{pair} values for each locus. The expected value of NG^{mepf}_{pair} (NG^{depf}_{pair} is analogous) is:

$$\overline{NG}^{mepf}_{pair} = \sum_{i=1}^{A}\sum_{j=1}^{A}\sum_{k=1}^{A}\sum_{m=1}^{A}(P_{li}\cdot P_{lj}\cdot P_{lk}\cdot P_{lm}\cdot NG^{mepf}_{ijkm}); \qquad \text{(A2.3)}$$

where $\overline{NG}^{mepf}_{pair}$ is the expected value of NG^{mepf}_{pair} for a given locus; and NG^{mepf}_{ijkm} is the value of NG^{mepf}_{pair} given the father's genotype consisting of alleles *i* and *j* and the mother's genotype consisting of alleles *k* and *m*, and is calculated as in Neff *et al*. (2000).

The expected value for multiple loci (when considered simultaneously) is the product of the expected NG^{mepf}_{pair} values for each locus. Finally, because the degree of polymorphism at a locus depends on both the number of alleles and their frequencies, it is useful to present an equation for the effective number of alleles:

$$A_e = \left(\sum_{a=1}^{A} P^2_{la}\right)^{-1}. \qquad \text{(A2.4)}$$

The effective number of alleles represents a locus with equivalent resolving power, but with A_e equally common alleles. The value of A_e can be used with Fig. 2 to determine the expected values of *NG* for the original locus having *A* alleles.