MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 115–128 doi: 10.1111/j.1755-0998.2009.02687.x

TECHNICAL ADVANCES

Parentage in natural populations: novel methods to detect parent-offspring pairs in large data sets

MARK R. CHRISTIE
Department of Zoology, Oregon State University, Corvallis, OR 97331, USA

Abstract

Parentage analysis in natural populations presents a valuable yet unique challenge because of large numbers of pairwise comparisons, marker set limitations and few sampled true parent–offspring pairs. These limitations can result in the incorrect assignment of false parent–offspring pairs that share alleles across multi-locus genotypes by chance alone. I first define a probability, Pr(δ), to estimate the expected number of false parent–offspring pairs within a data set. This probability can be used to determine whether one can accept all putative parent–offspring pairs with strict exclusion. I next define the probability Pr($\phi \mid \lambda$), which employs Bayes' theorem to determine the probability of a putative parent–offspring pair being false given the frequencies of shared alleles. This probability can be used to separate true parent–offspring pairs from false pairs that occur by chance when a data set lacks sufficient numbers of loci to accept all putative parent–offspring pairs. Finally, I propose a method to quantitatively determine how many loci to let mismatch for study-specific error rates and demonstrate that few data sets should need to allow more than two loci to mismatch. I test all theoretical predictions with simulated data and find that, first, Pr(δ) and Pr($\phi \mid \lambda$) have very low bias, and second, that power increases with lower sample sizes, uniform allele frequency distributions, and higher numbers of loci and alleles per locus. Comparisons of Pr($\phi \mid \lambda$) to strict exclusion and CERVUS demonstrate that this method may be most appropriate for large natural populations when supplemental data (e.g. genealogies, candidate parents) are absent.

Keywords: Bayes' theorem, dispersal, exclusion, gene-flow, genotyping error, parent–offspring pairs, paternity

Received 7 December 2008; revision accepted 27 February 2009

Introduction

Parentage analysis is a precise form of assignment testing (Manel *et al*. 2005) and can be particularly useful for detecting ecological and evolutionary patterns in systems with high levels of gene flow. Such systems have limited genetic differentiation, which severely restricts the utility of population-level assignment methods. Therefore, parentage analyses may allow for the inference of gene flow and dispersal at ecologically relevant timescales. A challenge to employing parentage analysis in natural populations is that large population sizes, variable dispersal distances and high rates of mortality may severely constrain the number of sampled parent–offspring pairs. These challenges are amplified in systems where patterns of dispersal are unobservable, such as the larval dispersal stage in the majority of marine fishes and invertebrates (Palumbi *et al*. 1997; Hixon *et al*. 2002; Leis 2006), where propagules are too small to track directly (but see Jones *et al*. 1999; Thorrold *et al*. 2006). In addition, because of a lack of pragmatic methods, long-distance dispersal events are often ignored or remain undetected in many species of plants (Nathan 2006), fungi (Kauserud *et al*. 2006) and animals that are cryptic or have complex life histories (Derycke *et al*. 2008). Large genotypic data sets may be used to uncover some of these enigmatic processes, and parentage analysis can be a powerful tool for the direct detection of patterns of dispersal and population connectivity.

Several studies have successfully employed parentage analyses to address questions of gene flow and dispersal. For example, parentage analysis has revealed patterns of

Correspondence: Mark R. Christie, Fax: (541) 737 0501; E-mail: christim@science.oregonstate.edu

dispersal in rodents (Telfer *et al.* 2003; Waser *et al.* 2006; Nutt 2008), insects (Tentelier *et al.* 2008) and fishes with dispersive larvae (Jones *et al.* 2005). Especially promising are recent attempts to estimate dispersal kernels with mean parent–offspring distances determined via parentage methods (Oddou-Muratorio *et al.* 2003; Robledo-Arnuncio & Garcia 2007). Moreover, parentage analyses could be coupled with population-level techniques in novel and effective ways. Direct estimates of parent–offspring dispersal could be incorporated as priors into Bayesian assignment methods or incorporated into a landscape genetics (Manel *et al.* 2003) or circuit-theory framework (McRae & Beier 2007). As parentage methods become more powerful and population-level methods increasingly detect fine-scale genetic structure, synergistic approaches hold great promise for accurate dispersal estimates.

The majority of studies using parentage methods to determine patterns of dispersal have relied upon likelihood-based approaches (Thompson 1975, 1976; Meagher 1986; Thompson & Meagher 1987). Several approaches have been suggested to evaluate the significance of likelihood ratios (Gerber *et al.* 2003; Anderson & Garza 2006) with CERVUS being the most commonly used program (Marshall *et al.* 1998; Kalinowski *et al.* 2007). Unfortunately, these methods of evaluating significance require estimates of demographic parameters often difficult or impossible to obtain from natural populations. The program CERVUS, for example, requires precise estimates of the number of candidate parents per offspring and the proportion of candidate parents sampled (Kalinowski *et al.* 2007). These parameters, along with a direct setting of the confidence level, serve to control type I and type II errors. However, in many cases, this process obfuscates parentage analyses because it remains unclear how sensitive CERVUS is to estimates of these parameters. Therefore, this approach may not be appropriate for many natural populations, when the probability of finding a parent is low and where reliable observational data are difficult to obtain.

For natural populations with few sampled parents, strict exclusion or kinship techniques are the preferred analytical approaches for parentage assignment (Jones & Ardren 2003). Kinship methods are restrictive because they determine only whether a data set has more related individuals than expected by chance (Queller *et al.* 2000), but often cannot identify which individuals those are. Strict exclusion, which is the process of excluding dyads through Mendelian incompatibility, is a powerful method. However, one must first determine whether their data set has enough polymorphic markers to minimize the occurrence of false pairs (i.e. adults that share an allele with a putative offspring by chance). As a consequence, many exclusion probabilities have been developed for a variety of applications. Some approaches focus on data sets where the genotypes of the mother and putative sire, or at least one parent, are available (Chakraborty *et al.* 1988; Jamieson & Taylor 1997), whereas other exclusion methods focus on excluding only a handful of candidate parents (Dodds *et al.* 1996). One exclusion probability that is appropriate for situations where neither parent is known was first described by Garber & Morris (1983) and later expressed in terms of homozygotes (Jamieson & Taylor 1997). Here, I show that this exclusion probability is biased when there are differences in allele frequencies between samples of adults and juveniles and recommend an unbiased alternative.

When applied correctly, exclusion is a powerful parentage method because it fully accounts for the uniqueness of the parent–offspring relationship (Milligan 2003) without any assumptions. It is this strength, however, that is often the greatest drawback to exclusion-based approaches because a genotyping error at a single locus can invalidate a true parent–offspring pair. In contrast to likelihood methods, it has proven difficult to incorporate genotyping error into exclusion-based methods. Thus, the majority of exclusion-based studies usually allow for a certain number of loci to mismatch (e.g. Vandeputte *et al.* 2006; McLean *et al.* 2008). This simply means that if a locus for a putative parent–offspring pair does not share an allele, then that locus is dropped for the analyses of that particular putative parent–offspring pair. Therefore, to fully account for genotyping error, it is necessary to start a project with a few more loci than the minimum required for sufficient exclusionary power. One major concern is deciding how many loci should be allowed to mismatch, and to date, this has largely been a subjective process (Hoffman & Amos 2005). If too many loci are allowed to mismatch, one runs the risk of falsely assigning parent–offspring pairs. If too few loci are allowed to mismatch, then one runs the risk of not identifying true parent–offspring pairs. Thus, I propose a quantitative approach to determine how many loci to allow to mismatch based upon study-specific estimates of genotyping error.

In this paper, I define the probability $\Pr(\delta)$, which is an unbiased exclusion probability that can be applied when parents are unknown. This probability can simply be multiplied by the total number of pairwise comparisons to estimate the number of false parent–offspring pairs within that data set. If a data set contains insufficient numbers of loci, such that it generates an unacceptable probability of containing false parent–offspring pairs, it is still possible to separate true from false parent–offspring pairs. To do so, I define a second probability, $\Pr(\phi \mid \lambda)$, that determines the probability of a putative parent–offspring pair being false given the frequencies of shared alleles. This novel approach allows researchers to identify true parent–offspring pairs when there is insufficient power for strict exclusion and, importantly, does

not require any estimates of demographic parameters. I then describe an approach to determine how many loci to let mismatch based upon study-specific error rates. Software to implement all analyses presented here are available at http://sites.google.com/site/parentagemethods/. In what follows, I first describe these methods and subsequently validate them by measuring bias in simulated data sets and by drawing comparisons between existing methods.

Methods

False parent–offspring pairs

Here, I describe the probability of false parent–offspring pairs occurring within a data set. This probability can determine whether the information content of one's data set is sufficient to accept all putative parent–offspring pairs with simple Mendelian incompatibility. This framework is developed assuming the use of co-dominant markers in diploid organisms. I also include a table that provides explicit definitions of terms used throughout this paper, as terminology varies across studies (Table 1).

The probability of a randomly selected dyad from a particular locus sharing an allele equals:

$$\Pr(Z) = \sum_{i=1}^{Na}\left(2z_{1i} - z_{1i}^2\right)\left(2z_{2i} - z_{2i}^2\right) - \sum_{i=1}^{Na-1}\sum_{g=i+1}^{Na}\left(2z_{1i}q_{1g}\right)\left(2z_{2i}q_{2g}\right) \quad (1)$$

where Na equals the total number of alleles at a locus, z_1 equals the allele frequency for allele i in the sample of adults and z_2 equals the allele frequency for allele i in the sample of juveniles. Thus, z_1^2 and z_2^2 equal the frequency of homozygotes containing allele i in samples of adults and juveniles, respectively, assuming Hardy–Weinberg Equilibrium (HWE). Alleles occurring in only one sample (i.e. adults or juveniles) will not be included in the above expression because the product equals zero. Notice that the expected number of homozygotes for an allele is subtracted from the total number of times the same allele occurs to prevent dyads that are homozygous for the same allele from being counted twice. Likewise, it is important to count only dyads that are heterozygous for the same alleles only once. Therefore, I subtract a double summation where q equals the frequencies of alleles (i+1):Na and where z_1q_1 and z_2q_2 are used to calculate the HWE-expected genotype frequencies of unique heterozygotes in samples of adults and juveniles respectively.

Under some circumstances, it may be desirable to use an equation that does not employ HWE estimates of genotype frequencies. One example would be if genotype frequency estimates have high accuracy yet do not conform to HWE expectations. The equation that does not assume HWE is:

$$\Pr(Z_G) = \sum_{i=1}^{Na}(2z_{1i} - zz_{1i})(2z_{2i} - zz_{2i}) - \sum_{i=1}^{Ng}(zq_{1i})(zq_{2i}) \quad (2)$$

where zz_1 and zz_2 equal the observed frequencies of homozygotes containing allele i in the samples of adults and juveniles, respectively, and zq_1 and zq_2 equal the observed frequencies of all unique heterozygotes, Ng, in the samples of adults and juveniles respectively.

To expand this approach to multiple loci, it is assumed throughout this paper that loci are in linkage equilibrium and are thus independent of one another. However, linked loci could be incorporated by explicitly accounting for the dependence between loci provided

Table 1 Definitions of terms used throughout the paper

Term	Definition
Adult	Any individual from a sample of sexually mature individuals
Parent	The true mother or father of an individual in a data set
Juvenile	Any individual from a sample of sexually immature individuals
Offspring	An individual that has a parent within the sample of adults
Dyad	Any pairwise comparison between an adult and a juvenile
False parent–offspring pair	A dyad that shares at least one allele at all loci by chance because of large sample sizes or insufficient numbers of loci or alleles per locus
True parent–offspring pair	A dyad that shares at least one allele at all loci because of direct Mendelian transmission
Putative parent–offspring pair	A dyad that shares at least one allele at all loci but that has yet to be assigned as a true or false parent–offspring pair
Type I error	A dyad that shares alleles across all loci by chance and is falsely determined to be a true parent–offspring pair
Type II error	A true parent–offspring pair that is not identified as such. Is often, along with type I error, expressed as a rate
Power	One minus the type II error rate. The proportion of true parent–offspring pairs that are correctly assigned

that estimates of recombination rates can be obtained (see methods in Thompson & Meagher 1998). If the assumption of linkage equilibrium is valid, it is possible to multiply probabilities across loci such that:

$$\Pr(\delta) = \prod_{i=1}^{L} \Pr(Z)_i \quad (3)$$

where L equals the total number of loci. To determine the approximate number of false parent–offspring pairs, *Fpairs*, for a given data set, $\Pr(\delta)$ should be multiplied by the total number of pairwise comparisons:

$$Fpairs = \Pr(\delta) \cdot n_1 \cdot n_2 \quad (4)$$

where n_1 equals the number of adults and n_2 equals the number of juveniles. It is important to keep in mind that this is a probability, and that variance due to sampling will cause slight deviations from this quantity. However, on average, these equations predict the number of false pairs very accurately (Fig. 1, Table 2). Notice that eqn 1 bears some similarity to the exclusion equations described in Jamieson & Taylor (1997). However, the exclusion equations presented by Jamieson and Taylor use allele frequencies from the combined samples of adults and juveniles, which results in positively biased estimates if there are only slight differences in allele frequencies between samples of adults and juveniles (Fig. 2, Table 3).

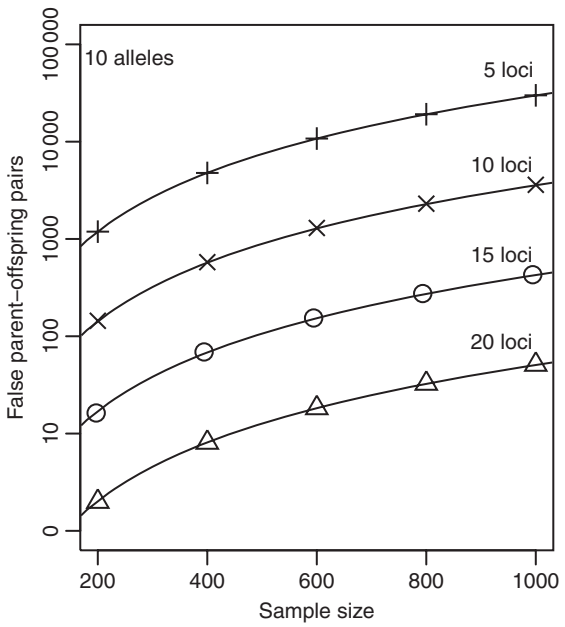


Fig. 1 Actual and predicted number of false parent–offspring pairs as measured by simulated data sets and as predicted by *Fpairs*. Each value was determined from 1000 simulated data sets with 10 alleles per locus. The actual number of false parent–offspring pairs are shown as + for data sets with 5 loci, × for data sets with 10 loci, ○ for data sets with 15 loci, and as △ for data sets with 20 loci. The black lines are the predicted values of the number of false parent–offspring pairs, *Fpairs*, with each line calculated using the same number of loci as the symbols that lay on it. Sample size equals the number of adults plus the number of juveniles, both of which are of equal size.

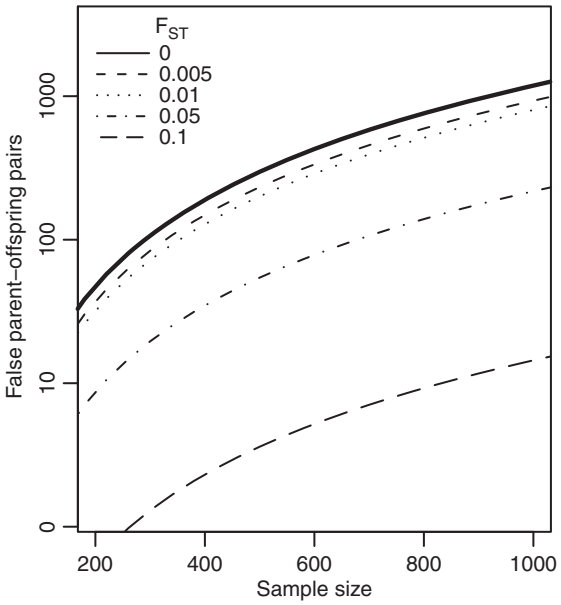


Fig. 2 Actual and predicted number of false parent–offspring pairs as measured by simulated data sets and as predicted by *Fpairs*. The lines represent the predicted values of the number of false parent–offspring pairs, *Fpairs*, from simulated data sets with 10 loci and 10 alleles per locus. Actual values were measured for combined sample sizes of 200, 400, 600, 800 and 1000 individuals. Actual values lay directly over the predicted values and symbols were omitted for clarity. F_{ST} between adults and juveniles was varied from 0 to 0.1. As F_{ST} increases, there is a decrease in the actual number of false parent–offspring pairs, which is matched precisely by *Fpairs*. The solid black line represents $F_{ST} = 0$, but also equals the predicted number of false parent–offspring pairs for methods that do not take into account the allele frequencies of adults and juveniles separately, regardless of the value of F_{ST}, and are thus positively biased.

The importance of minimizing the number of false parent–offspring pairs depends on the study, although the utility and accuracy of any parentage analysis obviously deteriorate as the number of false parent–offspring pairs increases. If the expected number of false parent–offspring pairs is negligible (i.e. near 0), then strict exclusion can be safely used. Here, the probability of any particular putative parent–offspring pair being false, when using strict exclusion, equals:

$$\Pr(\phi) = \frac{Fpairs}{N_P} \quad (5)$$

where N_P equals the observed number of putative parent–offspring pairs, which is simply calculated by

Table 2 Mean predicted number of false parent–offspring pairs and the bias between predicted and the actual number of false pairs. True pairs equal the number of true parent–offspring pairs, and NL, NA and N equal the number of loci, the number of alleles per locus and sample size respectively, where N equals the number of adults plus juveniles. Values were calculated from 1000 simulated data sets

True pairs	NL	NA	N	False pairs	Bias	$\sqrt{MSE}$
0	10	10	250	221.67	−0.0002760	0.00219
15	10	10	250	221.82	−0.0001316	0.00216
30	10	10	250	223.59	−0.0000362	0.00216
60	10	10	250	225.61	0.0002410	0.00221
0	10	15	250	57.01	−0.0001562	0.00222
15	10	15	250	57.53	−0.0000553	0.00220
30	10	15	250	57.56	−0.0000487	0.00217
60	10	15	250	57.31	−0.0000801	0.00216
0	15	10	250	28.59	−0.0000627	0.00223
15	15	10	250	29.92	0.0000353	0.00249
30	15	10	250	29.01	0.0000103	0.00228
60	15	10	250	28.35	−0.0000802	0.00228
0	15	10	500	108.54	−0.0000011	0.00220
15	15	10	500	108.53	−0.0000091	0.00224
30	15	10	500	108.67	0.0000057	0.00225
60	15	10	500	108.55	−0.0000058	0.00223
0	15	15	500	13.13	−0.0000038	0.00233
15	15	15	500	13.12	−0.0000047	0.00230
30	15	15	500	13.2	−0.0000014	0.00218
60	15	15	500	13.2	0.0000012	0.00224

summing the number of dyads that share at least one allele at all loci. N_P is also equal to the total number of false parent–offspring pairs plus the total number of true parent–offspring pairs. Because $\Pr(\phi)$ equals the probability of any putative parent–offspring pair being false, one should strive to minimize this value by employing many polymorphic loci. In addition, it may be useful to obtain an a priori estimate of $\Pr(\delta)$, decide upon an acceptable number of false pairs and solve for the maximum sample size for a particular marker set. Such a priori estimates can aid in determining whether more loci should be developed before performing parentage analyses.

Putative parent–offspring pairs

If the probability of type I error for strict exclusion is unacceptably high, such that it is unwise to accept all putative parent–offspring pairs as true pairs, it is often possible to determine whether some putative parent–offspring pairs are true pairs. This is achieved by calculating the probability of a putative parent–offspring pair being false given the frequencies of shared alleles, which using Bayes' theorem equals:

$$\Pr(\phi|\lambda) = \frac{\Pr(\lambda|\phi) \cdot \Pr(\phi)}{\Pr(\lambda)} \tag{6}$$

where $\Pr(\phi)$ equals the probability of a putative parent–offspring pair occurring by chance and $\Pr(\lambda)$ equals the probability of observing the shared alleles. $\Pr(\phi)$ is defined by eqn 5, yet we still need to define $\Pr(\lambda \mid \phi)$, the probability of observing the shared alleles given that the putative parent–offspring pair is false.

To calculate $\Pr(\lambda \mid \phi)$, one must first calculate a measure of the shared allele frequencies in a putative parent–offspring pair and second create a distribution of similar values generated from false pairs for comparison. It is important to note that it does not matter what measure of shared allele frequencies is used. Here, I employ an approach similar to eqn 1 to calculate an overall measure of shared allele frequencies, but one could just as easily use common likelihood methods (e.g. Thompson 1976), as the results would be identical. As before, each locus is treated independently. Thus the measure of shared allele frequencies employed here equals:

$$\Pr(\widetilde{Z}) = \sum_{i=1}^{\widetilde{Na}} \left(2z_{1i} - z_{1i}^2\right)\left(2z_{2i} - z_{2i}^2\right) - \sum_{i=1}^{\widetilde{Na}-1} \sum_{g=i+1}^{\widetilde{Na}} \left(2z_{1i}q_{1g}\right)\left(2z_{2i}q_{2g}\right) \tag{7}$$

where all symbols are the same as eqn 1, except that $\widetilde{Na}$ equals the number of alleles, including the shared allele, that occur with a frequency less than or equal to that of the shared allele. Because this approach only examines the frequency of the one shared allele in accordance with Mendelian inheritance, if a putative parent–offspring pair happens to be heterozygous for the same alleles, it is appropriate to employ the rarer of the two alleles in the above framework. This probability can once again be combined across all loci, assuming linkage equilibrium, such that:

$$\Pr(\widetilde{\delta}) = \prod_{i=1}^{L} \Pr(\widetilde{Z})_i \tag{8}$$

where this equation represents the probability of observing a dyad, not a putative parent–offspring pair, that shares equally or less common alleles.

To determine $\Pr(\lambda \mid \phi)$, a distribution of $\Pr(\widetilde{\delta})$ from false parent–offspring pairs must be created. This is achieved by creating data sets (hereafter referred to as null sets) with the same allele frequencies, sample sizes and number of loci as the real data set of interest. These null sets contain no true parent–offspring pairs; thus, all simulated adults and juveniles that share at least one allele across all loci do so by chance alone (i.e. all putative pairs are false parent–offspring pairs). For every null data set, $\Pr(\widetilde{\delta})$ is calculated for every false parent–offspring

pair. These values are used to create a distribution of false parent–offspring pairs. To reduce bias, at least 10 000 individual false $\Pr(\widetilde{\delta})$ values should be generated from a minimum of 100 null sets, which was found to be more than sufficient under all conditions tested. Notice that the number of values used to create this distribution does not come from any assumptions about the data. It is only necessary to ensure that this distribution is representative of the true distribution of false pairs, with more calculated values creating a more accurate description of the distribution. The mean and variance of this distribution depend upon the power of the data set used to create it. Figure 3a shows an example of such a distribution created from 10 000 false pairs.

To calculate $\Pr(\lambda \mid \phi)$, the value of $\Pr(\widetilde{\delta})$ for the putative pair under consideration, $\Pr(\widetilde{\delta})_i$, is compared with the distribution of values generated by false parent–offspring pairs, $\Pr(\widetilde{\delta})_F$, such that:

$$\Pr(\lambda|\phi) = \frac{N\left(\Pr(\widetilde{\delta})_F \leq \Pr(\widetilde{\delta})_i\right)}{N(\Pr(\widetilde{\delta})_F)} \tag{9}$$

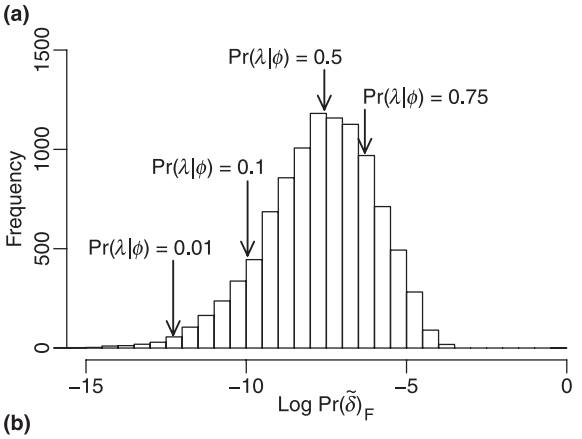
where the numerator equals the total number of false values, $\Pr(\widetilde{\delta})_F$, generated from null sets, that are less than or equal to $\Pr(\widetilde{\delta})_i$ and the denominator equals the total number of false $\Pr(\widetilde{\delta})_F$ values used to create the distribution. For example, if 100 $\Pr(\widetilde{\delta})_F$ values out of a distribution of 10 000 were found to be less than or equal to $\Pr(\widetilde{\delta})_i$, then $\Pr(\lambda \mid \phi)$ would equal 0.01 (Fig. 3a).

The remaining probability needed to satisfy eqn 6 is $\Pr(\lambda)$, the probability of observing the shared alleles. This is obtained by noticing that Bayes' theorem is often restated (see Sokal & Rohlf 1995 for a general treatment; Carlin & Louis 2000 for a detailed treatment):

$$\Pr(\phi|\lambda) = \frac{\Pr(\lambda|\phi) \cdot \Pr(\phi)}{\Pr(\lambda|\phi) \cdot \Pr(\phi) + \Pr(\lambda|\phi^C) \cdot \Pr(\phi^C)} \tag{10}$$

where $\Pr(\phi^C)$ equals the complement of $\Pr(\phi)$ and where $\Pr(\lambda \mid \phi^C)$ equals unity. This is because $\Pr(\lambda \mid \phi^C)$ equals the probability of observing the shared alleles given that the putative parent–offspring pair is true. There should be no reason for a true parent–offspring pair to be constrained to any particular set of alleles, thus this value should equal unity unless there is selection for or against alleles. Notice that if $\Pr(\lambda \mid \phi)$ equals one, meaning that the putative parent–offspring pair shares the most common alleles at all loci, then $\Pr(\lambda \mid \phi) = \Pr(\phi)$. In addition, it is clear that if rarer alleles are shared, then it becomes less likely that a putative parent–offspring pair is false (Fig. 3b).

As with most statistical methods, an arbitrary cut-off value can be decided upon a priori. Choosing a cut-off value is largely a matter of convenience and may depend on the goals of the study. The interpretation of $\Pr(\phi \mid \lambda)$ is straightforward: it represents the probability of a putative parent–offspring pair being false given the frequencies of shared alleles. It is important to recognize that a large probability does not mean that a dyad is a false parent–offspring pair, but rather that more data are needed in the form of additional loci.



(b)

	Pr(λ|φ)				
Pr(φ)	0.01	0.1	0.25	0.5	0.75
0.90	0.083	0.474	0.692	0.818	0.871
0.80	0.039	0.286	0.500	0.667	0.750
0.60	0.015	0.130	0.273	0.429	0.529
0.40	0.007	0.063	0.143	0.250	0.333
0.20	0.003	0.024	0.059	0.111	0.158
0.10	0.001	0.011	0.027	0.053	0.077
0.05	0.001	0.005	0.013	0.026	0.038
0.01	0.000	0.001	0.003	0.005	0.008

Fig. 3 (a) Distribution of $\Pr(\widetilde{\delta})_F$ values created from 10 000 false parent–offspring pairs generated from simulated null data sets with 15 loci, 15 alleles per locus, 500 individuals and no true parent–offspring pairs. Various probabilities of observing a set of shared alleles given that a putative parent–offspring pair is false, $\Pr(\lambda \mid \phi)$, are indicated with arrows. (b) A table of $\Pr(\phi \mid \lambda)$ values, the probability of a putative parent–offspring pair being false given the frequencies of alleles that it shares. Notice that the value of $\Pr(\phi \mid \lambda)$ is dictated by two values, $\Pr(\phi)$ and $\Pr(\lambda \mid \phi)$. Thus, $\Pr(\phi \mid \lambda)$ will be small if the putative parent–offspring pair shares rare alleles (i.e. $\Pr(\lambda \mid \phi)$ is small), or if $\Pr(\phi)$ is small, or both.

Genotyping error

If a putative parent–offspring pair does not share an allele at a singe locus, then $\Pr(\phi \mid \lambda)$ remains undefined. To allow for genotyping errors, null alleles and mutations, it is necessary to quantitatively estimate how many loci should be allowed to mismatch based upon the study-specific error rate. First, one must perform a second independent analysis on a subset of genotyped samples across all loci. The study-specific error rate, ε, is then defined as the quotient of the number of alleles

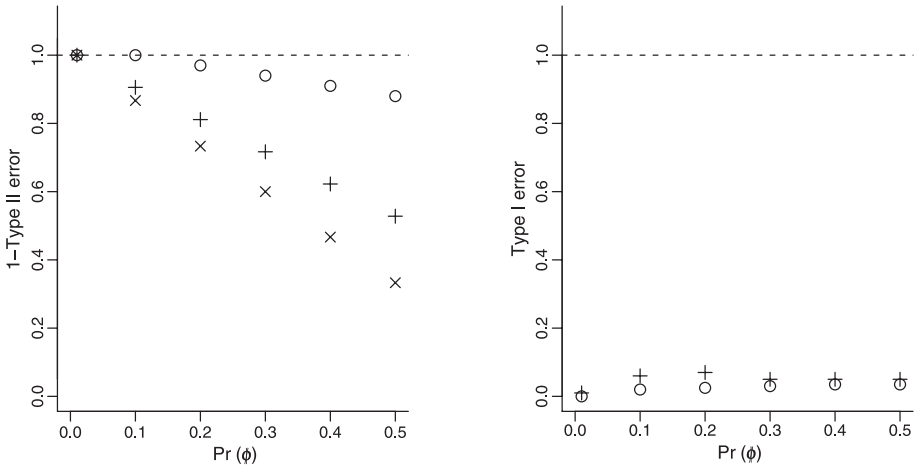


Fig. 4 Comparisons of $\Pr(\phi \mid \lambda)$ to strict exclusion and CERVUS across varied levels of $\Pr(\phi)$. The left plot examines the proportion of true parent–offspring pairs that are correctly identified as true parent–offspring pairs, 1 minus type II error, which is equivalent to the power. The right plot examines the proportion of false parent–offspring pairs that are incorrectly assigned as true parent–offspring pairs, type I error. Results for strict exclusion are shown with a dashed line. Results for $\Pr(\phi \mid \lambda)$ are shown with ○. Results for CERVUS are shown with a + for data sets where the number of candidate parents was set to 500, and with an × where the number of candidate parents was set to 1000. Type I error for CERVUS was only reported for data sets where the number of candidate parents was set to 500, although the results were nearly identical for the data sets with 1000 candidate parents.

that differ after the second analysis to the total number of alleles compared (*sensu* Bonin *et al.* 2004). To estimate quantitatively the number of loci that should be allowed to mismatch, one must first determine the probability of observing at least one error in a multi-locus genotype. I use a simplification of Bonin *et al.* (2004) formula:

$$P = 1 - (1 - \varepsilon)^{2L} \tag{11}$$

where L is equal to the total number of loci employed in the study. This probability comes from solving the binomial for the proportion of multi-locus genotypes with no errors and subtracting the result from unity to account for all errors. Alternatively, one can solve for the proportion of multi-locus genotypes that has exactly i errors:

$$P_i = \binom{2L}{i} (\varepsilon)^i (1 - \varepsilon)^{2L-i} \tag{12}$$

I extend this probability to determine the probability of observing at least one error in a multi-locus pairwise comparison. This probability is equal to:

$$P' = 2P - P^2 \tag{13}$$

where P' equals the proportion of pairwise comparisons (i.e. dyads) that will have at least one error at a locus. Notice that there is no solution for L, without the use of imaginary numbers. However, one can iteratively determine the proportion of dyads that would have at least one error given a number of mismatching loci:

$$P'_i = 2\left(P - \sum_{i=1}^{M} P_i\right) - \left(P - \sum_{i=1}^{M} P_i\right)^2 \tag{14}$$

where M equals the number of loci allowed to mismatch and must be an integer greater than 0. Thus P'_i equals the number of dyads with at least one error given that M loci are allowed to mismatch. Not all errors will cause a mismatch because the majority of dyads will not be parent–offspring pairs and in addition, the majority of positions where an error occurs will not cause a mismatch. Thus choosing a cut-off value for P'_i is somewhat subjective and should be reported along with the number of loci allowed to mismatch. As a general rule of thumb, a P'_i between 0.05 and 0.1 will eliminate the majority of putative pairs with a mismatch-causing error. The advantage to this method becomes quickly apparent when one notices how quickly the error rate drops by allowing a single locus to mismatch. Therefore, while a P'_i near 0.05 may be conservative, in most cases it is approached by simply allowing one to two loci to mismatch (Fig. 5). Notice that this approach can also be applied to methods that determine the probability of identity among genotypes and that one can additionally account for null alleles, missing data and mutation simply by adding estimates of those rates to ε.

Validation

The above methodology was tested with simulated data sets to determine whether the theoretical predictions

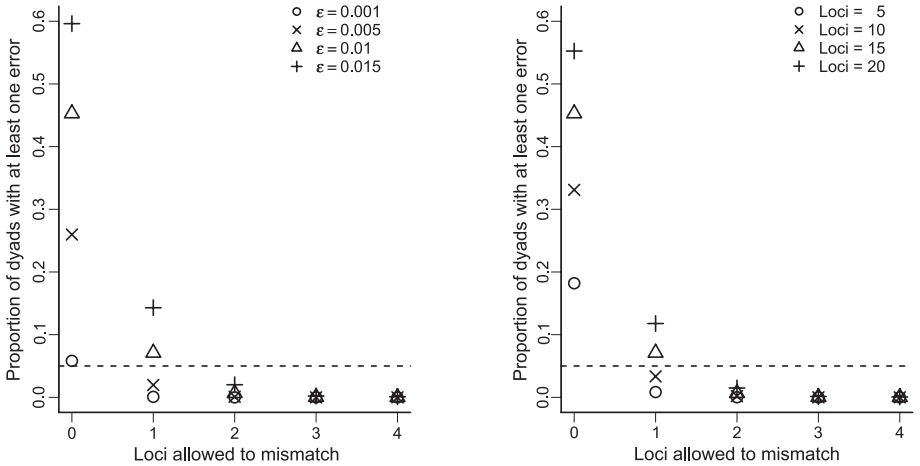


Fig. 5 The proportion of dyads with at least one error as a result of allowing different numbers of loci to mismatch. The proportion of dyads with at least one error was calculated with eqn 14 and is dependent upon the error rate and number of loci. The plot on the left examines the effect of varied error rates (ε), while holding the number of loci at 15. The plot on the right examines the effect of varied numbers of loci, while holding the error rate at 0.01.

matched actual occurrences of false parent–offspring pairs and to compare with existing methods. All data simulation and probability calculations were implemented in R version 2.5.1 (R Development Core Team 2007). Simulated data sets were created using a set of alleles whose frequencies were determined by the equation:

$$z = \frac{i}{(Na+1)-i} \cdot \sum_{i=1}^{Na} \frac{i}{(Na+1)-i} \qquad (15)$$

where Na equals the total number of alleles and i equals allele i in the set 1:Na. This distribution is fairly conservative as it results in several fairly common alleles (Bernatchez & Duchesne 2000). Allele frequency distributions had no effect on the precision and accuracy of these methods, but in all cases, a uniform distribution (i.e. equal allele frequencies) resulted in the greatest power. Once the population allele frequencies were determined, 100 000 genotypes were created in accordance with HWE. This pool of genotypes was randomly sampled and placed into a group of either adults or juveniles. This process was repeated for each locus. True parent–offspring pairs were created by randomly sampling one individual from both the sample of adults and the sample of juveniles. The two individuals were aligned, locus by locus, and at each locus, a randomly chosen allele was copied from the adult to the offspring. This procedure was executed regardless of whether or not the pair already shared an allele at that locus, which simulated the occasional, but realistic, occurrence of dyads being homozygous or heterozygous for the same alleles. This procedure also has the benefit of making the distribution of shared alleles equal to that of the overall sample distribution, which is expected with a random sample.

I first tested whether the theoretical predicted number of false pairs, as calculated by Pr(δ), matched the actual number of false parent–offspring pairs. Simulated data sets were created with varied numbers of loci, sample sizes, alleles and true parent–offspring pairs. All simulated data sets had equal sample sizes of adults and juveniles, which maximizes the number of pairwise comparisons. One thousand simulated data sets were created for each combination of variables. For each data set, theoretical estimates of the expected number of false parent–offspring pairs were calculated using eqn 4 and compared with the actual number of false pairs observed. The bias, root mean square error and variance of the predicted number of false pairs were calculated from all 1000 simulated data sets.

I next tested whether Pr(δ) was less biased than the multi-locus approach presented by Jamieson & Taylor (1997), represented in their paper with equations two through four and hereafter denoted as P. I manipulated F_{ST} between the samples of adults and juveniles from 0 to 0.1. Here, and only for this section, a uniform allele frequency distribution was employed to simplify the creation of different F_{ST} values. Because identical F_{ST} values can be created with different combinations of allele frequencies, I adjusted the frequencies such that the mean allele frequencies for adults and juveniles always equalled the starting allele frequencies when F_{ST} equalled zero. For example, in the two-allele case, F_{ST} equalled 0 when both alleles were set to 0.5. However, if one allele was increased to 0.6 in adults, then the same allele was decreased to 0.4 in juveniles so that the mean allele frequency remained 0.5. This process creates a standardized approach to creating F_{ST} values and highlights the differences between the two equations. The differences in

estimates were both plotted and presented in terms of bias. I additionally report a standardized bias, which equals the bias divided by the total number of false parent–offspring pairs when F_{ST} equals zero. This was to demonstrate that although the bias appears to decrease with increasing power, this is only an artefact of there being far fewer false parent–offspring pairs, and that proportionally the bias is much greater in data sets with greater power.

I first validated $\Pr(\phi \mid \lambda)$ by ensuring that the chosen type I error rate matched the actual type I error rate. To do this, I created simulated data sets with varying probabilities of a putative parent–offspring pair being false, $\Pr(\phi)$. All data sets were tested with 1, 30 and 60 true parent–offspring pairs and 1000 simulated data sets were created for each combination of values. The type I error rate, α, was set at 0.001, 0.01, 0.05 and 0.09 for each data set. Bias was measured by comparing the difference between the observed error rate and the set error rate, over all simulated data sets. For all analyses, the larger of 100 null sets or 10 000 $\Pr(\tilde{\delta})_F$ values was used for the calculation of $\Pr(\lambda \mid \phi)$.

I next compared $\Pr(\phi \mid \lambda)$ with CERVUS v. 3.0, again with data sets created with varied $\Pr(\phi)$ values. To accomplish this, sample size, numbers of loci and alleles per locus, and the numbers of true parent–offspring pairs were varied. Both the proportion of true parent–offspring pairs correctly assigned as true pairs and the proportion of false parent-offspring pairs incorrectly assigned as true pairs were recorded for both methods. Direct comparisons to CERVUS are difficult to make because CERVUS requires the estimates of two parameters that $\Pr(\phi \mid \lambda)$ does not require: the number of candidate parents and the proportion of candidate parents sampled. For all comparisons, the number of candidate parents was set to either 500 or 1000 and the strict confidence level of 95% was used. The proportion of candidate parents was set to the true parameter value. This is equitable because comparisons with larger numbers of candidate parents, and inaccurate estimates of the proportion of candidate parents sampled resulted in poor performance by CERVUS (see Fig. 4 of Marshall *et al.* 1998). Putative parent–offspring pairs were accepted as true parent–offspring pairs if $\Pr(\phi \mid \lambda)$ was less than or equal to 0.05.

Finally, I test eqn 14, which predicts the proportion of dyads with at least one error for a given error rate and number of loci allowed to mismatch. I first demonstrate how the proportion of dyads with at least one error is affected by varying error rates and numbers of loci. I next created simulated data sets with 15 and 20 loci, each with error rates of 0.01 and 0.015. The number of loci allowed to mismatch was varied from one to four, and the proportion of true parent–offspring pairs correctly assigned was recorded.

Results

The theoretical predictions for the number of false parent–offspring pairs, as determined by eqn 4, match very closely to the observed number of false parent–offspring pairs from the simulated data sets (Fig. 1, Table 2). Not surprisingly, the number of false parent–offspring pairs increases as sample size increases. In addition, there are fewer false parent–offspring pairs in data sets with more loci. The rate of increase in false parent–offspring pairs is identical in data sets with different numbers of loci, but identical allele frequency distributions. Table 2 demonstrates that increasing numbers of true parent–offspring pairs do not affect the predictive performance of $\Pr(\delta)$. The bias is slightly larger in the data sets with smaller sample size. Numbers of loci or alleles do not influence bias, provided that the sample size is large enough to accurately estimate allele frequencies. Overall, the methods presented here predict the actual number of false parent–offspring pairs with high accuracy and precision.

$\Pr(\delta)$ is unbiased when there are differences in allele frequencies between the samples of adults and juveniles (Table 3). However, the approach employed by Jamieson & Taylor (1997) reveal that even small differences in allele frequencies between adults and juveniles can result in a large overestimation in the number of false parent–offspring pairs (Fig. 2, Table 3). The bias in their method increases with increasing genetic differentiation, whereas $\Pr(\delta)$ remains unbiased regardless of the level of genetic differentiation. In addition, while the bias in their approach decreases with increasing number of alleles and loci, the standardized bias increases with increasing numbers of alleles and loci. This demonstrates that although the bias appears to decrease with increasing power, this is only an artefact of there being far fewer false parent–offspring pairs, and that proportionally the bias is much greater in data sets with greater exclusionary power.

The bias in type I error for $\Pr(\phi \mid \lambda)$ was very low across all tested levels of α, regardless of the number of true parent–offspring pairs or the value of $\Pr(\phi)$ (Table 4). The bias in type I error does not appear to follow any trends with number of true pairs, or type of data set. Moreover, the bias was negative in all cases meaning that the actual type I error rate was minutely smaller than the set type I error rate, thus making this approach, if anything, conservative. Overall, the very low rates of bias suggest that $\Pr(\phi \mid \lambda)$ can be used confidently to determine the probability of a putative parent–offspring pair sharing alleles by chance. The comparisons of $\Pr(\phi \mid \lambda)$ to strict exclusion demonstrate the utility of each method (Fig. 4). Not surprisingly, strict exclusion had a lower type II error than $\Pr(\phi \mid \lambda)$, correctly identifying all true parent–offspring pairs. In fact, with no genotyping error,

Table 3 Bias between the predicted exclusion probability and the actual exclusion probability used to calculate the expected number of false parent–offspring pairs. Two exclusion probabilities were compared: Jamieson & Taylor's (1997) eqn 2, denoted as P, and Pr(δ). For P, two measures of bias are employed: the absolute bias, Bias P, which equals the predicted exclusion probability minus the actual exclusion probability, and the standardized bias, Bias^S P, which is the bias divided by P when F_{ST} equals 0. This measure was used to demonstrate that although the bias can be very small, this is only an artefact of the exclusion probability being very small

Alleles	Loci	FST	Bias P	Bias^S P	Bias Pr(δ)
2	10	0	0	0	0
2	10	0.005	0.011	0.041	0
2	10	0.01	0.022	0.083	0
2	10	0.05	0.098	0.374	0
2	10	0.1	0.160	0.607	0
2	15	0	0	0	0
2	15	0.005	0.008	0.061	0
2	15	0.01	0.016	0.121	0
2	15	0.05	0.068	0.505	0
2	15	0.1	0.102	0.754	0
10	10	0	0	0	0
10	10	0.005	8.69×10^{-6}	0.386	0
10	10	0.01	1.23×10^{-5}	0.544	0
10	10	0.05	2.18×10^{-5}	0.966	0
10	10	0.1	2.25×10^{-5}	1.000	0
10	15	0	0	0	0
10	15	0.005	5.55×10^{-8}	0.519	0
10	15	0.01	7.41×10^{-8}	0.692	0
10	15	0.05	1.06×10^{-7}	0.994	0
10	15	0.1	1.07×10^{-7}	1.000	0

Table 4 Bias between the set type I error rate and the observed type I error rate for Pr($\phi \mid \lambda$). The type I error rate, α, was set at 0.001, 0.01, 0.05 and 0.09 for each data set. True pairs equals the number of true parent–offspring pairs, and Pr(ϕ) equals the probability of a putative parent–offspring pair being false and decreases with increasing number of alleles and loci or decreasing sample size (see text for details). Each value was calculated from 1000 simulated data sets

		α			
Pr(ϕ)	True pairs	0.001	0.01	0.05	0.09
0.9	1	−0.00065	−0.00069	−0.00071	−0.00075
0.9	30	−0.00061	−0.00066	−0.00073	−0.00075
0.9	60	−0.00062	−0.00064	−0.00069	−0.00071
0.5	1	−0.00052	−0.00058	−0.00061	−0.00065
0.5	30	−0.0005	−0.00053	−0.00057	−0.00059
0.5	60	−0.00059	−0.00062	−0.00068	−0.00072
0.25	1	−0.00043	−0.00047	−0.00049	−0.00053
0.25	30	−0.00042	−0.00045	−0.00052	−0.00057
0.25	60	−0.00038	−0.00046	−0.00048	−0.00051
0.1	1	−0.00028	−0.00029	−0.00032	−0.00057
0.1	30	−0.00013	−0.00019	−0.00023	−0.00048
0.1	60	−0.00024	−0.00036	−0.00038	−0.00053

strict exclusion has a type II error rate of 0. However, there is no mechanism for strict exclusion to determine the difference between real and false parent–offspring pairs, such that its type I error rate always equals unity. Thus, if there are any false parent–offspring pairs in the data sets, they will be assigned as true pairs. This highlights the importance of using strict exclusion only after sufficient power has been quantitatively determined (e.g. eqn 5). The observed patterns of type II error for Pr($\phi \mid \lambda$) were as expected. As Pr(ϕ) increases, the proportion of true parent–offspring pairs that were detected decreases.

Comparisons with CERVUS were also informative (Fig. 4). The proportion of true parent–offspring pairs successfully assigned was lower for CERVUS than for Pr($\phi \mid \lambda$), except when Pr(ϕ) equalled 0, where they were equivalent. Not surprisingly, the proportion of parents successfully assigned with CERVUS was lower when the number of candidate parents was set to 1000 as opposed to 500, which highlights the problems of using CERVUS when there are potentially large numbers of candidate parents. However, CERVUS performs as well or better than Pr($\phi \mid \lambda$) when the number of candidate parents is low or when the number of candidate parents is close to the true number of parents in the sample (data not shown). Here, we used the parameter value for the proportion of candidate parents sampled, however, inaccurate estimates of this parameter can have large effects on both the type I and type II error. For both methods, the type I error was within acceptable limits, though it appeared to be slightly lower for Pr($\phi \mid \lambda$).

Lastly, I examined the effects of genotyping error. All data sets with greater numbers of loci and higher error rates had a larger proportion of dyads with at least one error (Fig. 5). The proportional rate of decrease in dyads with at least one error was much greater for data sets with more loci and higher error rates. This pattern is reflected in the simulated data sets, where data sets with more loci and higher error rates had a greater increase in power by allowing loci to mismatch (Fig. 6). The data sets with 15 loci and an error rate of 0.01 only required one locus to mismatch before the proportion of true parent–offspring pairs correctly assigned equalled 1. All other values required two loci to mismatch before the proportion of true parent-offspring pairs correctly assigned equalled 1. Despite the relatively high error rates and number of loci, no data set required more than two loci to mismatch before all of the true parent–offspring pairs were correctly identified.

Discussion

This paper introduces novel approaches for determining parentage in natural populations. The theoretical predictions of the number of false parent–offspring pairs

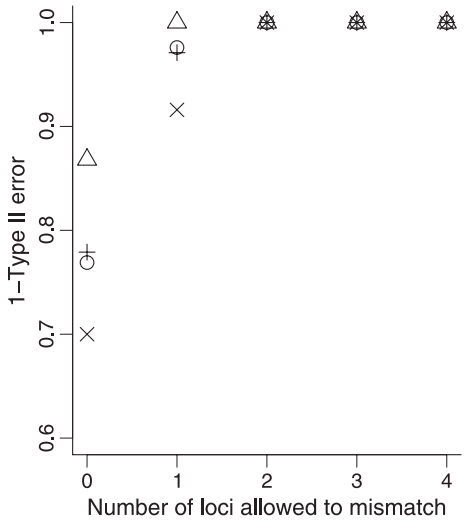


Fig. 6 The proportion of true parent–offspring pairs correctly assigned, 1 minus type II error (or power), as a function of the number of loci allowed to mismatch. Simulated data sets were created as follows: △ 15 loci, 0.01 error rate, + 15 loci, 0.015 error rate, ○ 20 loci, 0.01 error rate, × 20 loci, 0.015 error rate. Notice that even for a high error rate and relatively large numbers of loci, allowing for two loci to mismatch results in the correct assignment of all true parent–offspring pairs.

calculated with $\Pr(\delta)$ are matched very closely by the actual number of false parent–offspring pairs from simulated data sets. $\Pr(\delta)$ remains unbiased when there are differences in allele frequencies between adults and juveniles, to which prior methods are susceptible. There may still be occasions, however, when it is better to use allele frequencies from the combined sample of adults and juveniles, such as with small sample sizes or samples with inaccurate allele frequency estimates. It is worthwhile noting that even data sets with 20 loci had false parent–offspring pairs across all tested sample sizes suggesting that many studies employing strict exclusion may be plagued by false parent–offspring pairs. In fact, this highlights the need for any study employing Mendelian incompatibility to report some measure of exclusionary power. As the theoretical predictions of $\Pr(\delta)$ and the simulated data match well, and with little bias, this approach can be used confidently to determine how many false parent–offspring pairs are likely to exist in large data sets from natural populations.

Conveniently, one can conduct a priori analyses for a given marker set to determine the maximum sample size before type I error becomes problematic. Alternatively, with a simple rearrangement of eqn 4, one can determine the exclusionary power required for a desired sample size and number of false pairs. This approach is strongly recommended for any project in its infancy. However, in many cases, the data are already collected or there are only a limited number of loci available, at which point $\Pr(\phi \mid \lambda)$ should be used.

The analyses presented here demonstrate that $\Pr(\phi \mid \lambda)$ performs well for identifying true parent–offspring pairs without compromising the type I error. The importance of this result is particularly noticeable when one considers the type I error rate (i.e. the proportion of false parent–offspring pairs identified as true parent–offspring pairs) for strict exclusion equals unity. These results also show that, even in data sets where there is a fairly high ratio of false pairs to putative pairs, $\Pr(\phi \mid \lambda)$ is able to correctly identify the majority of true parent–offspring pairs. This result is encouraging because it indicates that this approach works well with data sets that do not have quite enough power to employ strict exclusion. It is important to point out that the average shared value of allele frequencies among true parent–offspring pairs is often much less than the average value among false pairs. This is because, as false parent–offspring pairs arise in data sets, they initially share the most common alleles. As false parent–offspring pairs become more prevalent (e.g. with increasing sample size), they gradually begin to share rarer alleles. It is this pattern, in part, that makes $\Pr(\phi \mid \lambda)$ a powerful method because it exploits the subtle differences in allele frequencies that exist between some true and false parent–offspring pairs. Furthermore, the intuitive results from Fig. 3, suggest that $\Pr(\phi \mid \lambda)$ may still be used to discern a few true parent–offspring pairs from data sets with very large numbers of false parent–offspring pairs, provided that the true parent–offspring pairs share rare alleles.

The methods presented here also have some distinct advantages over other commonly used methods, such as the likelihood approaches as implemented by CERVUS (Kalinowski *et al*. 2007), because no estimates of the number of candidate parents or the proportion of candidate parents sampled are needed. The results from data sets analysed by CERVUS demonstrate that power decreases rapidly for large numbers of candidate parents. Nonetheless, CERVUS works remarkably well for its intended application, and when many of the putative parents are sampled and accurate estimates of the required parameters are available, CERVUS is highly recommended. However, as is the case in many natural populations, when the number of true parent–offspring pairs within a large sample is low and the numbers of candidate parents are large, alternative approaches are required. Thus, new methods, such as the probability $\Pr(\phi \mid \lambda)$ introduced here, are needed to detect parent–offspring pairs in large data sets collected from natural populations with little to no genealogical information.

It is important to determine the effects of genotyping error for any parentage method, as even low levels of genotyping error may play a significant role in increasing

type II error for large data sets (Slate *et al.* 2000). However, the majority of errors will not cause a mismatch because: (1) most of the dyads are not true parent–offspring pairs and (2) most positions for an error to occur will not cause a mismatch. I demonstrate here that error has the largest effect on power when no loci are allowed to mismatch. However, the proportion of dyads with at least one error drops rapidly when just one locus is allowed to mismatch. For all the combinations of loci and error rates tested here, no data sets had detectable type II error after two loci were allowed to mismatch (exclusionary power was not an issue), which suggests that many studies may need only to allow one or two loci to mismatch. Interestingly, the initial rate of decrease in the proportion of dyads with a genotyping error is greater for data sets with more loci or higher error rates, meaning that these types of data sets have the most to gain by allowing loci to mismatch. Nevertheless, sufficient increases in power can be obtained even for data sets with 10 loci and an average error rate of 0.01. The methods presented here could be improved by determining the proportion of true parent–offspring pairs, as opposed to dyads, with at least one error. This challenging task would involve determining how often an error would occur between shared alleles after taking into account the allelic state of the individuals. As it stands, a cut-off level for the proportion of dyads with at least one genotyping error of between 0.05 and 0.1 should result in the maximum increase in power provided that there are sufficient numbers of loci.

With the rapid increases in DNA-based technology, an overwhelming number of markers may soon become available. However, given a set of equally polymorphic loci, one can see that the each additional locus provides diminishing returns (e.g. see eqn 3). The possible waste of time, money and effort associated with negligible gains in exclusionary power can again be avoided by performing a priori analyses for given marker sets and study designs. Moreover, employing too many loci may be counter productive because of increases in the study-wide error rate. At some point, diminishing returns will be reached between the addition of loci and the number of loci needed to mismatch to accommodate error. In light of recent trends towards employing large numbers of markers, I reiterate the importance of quantitatively determining the number of loci to allow to mismatch to avoid dropping loci that truly invalidate a putative parent–offspring relationship.

The parentage methods presented here suggest several promising avenues for further investigation. These methods could benefit from further testing and refinement under conditions that are rare, but present in certain natural populations. It is possible, for example, that highly skewed reproductive success or large sampling variances could create bias. It would also be useful to examine the effects of inbreeding and population substructure within samples of adults and juveniles as was recently performed for relatedness measures (Anderson & Weir 2007). In addition, where the presence of first-degree relatives other than parents and offspring may be an issue, I advise calculating the probability of related individuals sharing alleles at all loci for a given marker set and, if necessary, adding more loci (see Blouin 2003 and references therein). Finally, for data sets with very low numbers of false parent–offspring pairs (e.g. <1) or data sets with very large numbers of false parent–offspring pairs (e.g. >10 000), the simulations used to calculate $\Pr(\lambda \mid \phi)$ can be time consuming, such that alternative approaches for calculating $\Pr(\lambda \mid \phi)$ may be more efficient than the simulation-based procedure presented here.

The fields of parentage and kinship analysis have been growing rapidly since the development of hypervariable markers (Blouin 2003; Jones & Ardren 2003), yet there is a growing need to apply these approaches to large populations to uncover patterns of dispersal and gene flow. Many organisms have propagules that are difficult to track directly, and patterns of dispersal in these systems are not well understood. Parentage provides an approach that is analogous to mark–recapture studies and therefore may be especially useful in describing gene flow and dispersal at shorter time scales (e.g. among cohorts, years, or seasons) and among populations with little genetic differentiation. Moreover, with species that are relatively long-lived, one could construct a dynamic 'library' of potential parents with which to compare putative offspring year after year. Samples collected over subsequent years may provide clearer insights into patterns of dispersal. The direct dispersal information gained from parentage analysis can also be used to inform population-level analyses of dispersal and may be particularly useful when incorporated into population genetics (Manel *et al.* 2003; McRae & Beier 2007) or coupled with ecological or remote-sensing data to complement, validate and enhance nongenetic approaches for estimating dispersal. Parent–offspring information could also be coupled with population density estimates and isolation-by-distance analyses to provide more accurate estimates of species dispersal kernels (Rousset 1997). The theory and methods presented here have been developed in an effort to expand our capabilities of assessing gene flow and dispersal in large natural populations. Of course, these methods are also suited to answer a broad array of questions, such as determining associations between phenotypes or domestication and fitness (e.g. DeWoody 2005; Araki *et al.* 2007). Thus, parentage analysis in natural populations is vital for the advancement of both ecology (e.g. appropriate reserve design and spacing,

description of meta-population structure) and evolution (e.g. gene-flow estimation, selection, speciation).

Acknowledgements

This paper is a chapter from my doctoral dissertation and I thank my graduate committee for their reviews and support: Mark Hixon (chair), Steve Arnold, Michael Banks and Mike Blouin. I also wish to thank Hitoshi Araki, Mark Albins, Darren Johnson, Catherine Searle, Jacob Tennessen, the editors, and four anonymous reviewers for helpful comments that greatly improved this manuscript. This work was supported by grants from Conservation International and the National Science Foundation (05-50709) to M.A. Hixon.

References

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–2582.

Anderson AD, Weir BS (2007) A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, **176**, 421–440.

Araki H, Cooper B, Blouin MS (2007) Genetic effects of captive breeding cause a rapid, cumulative fitness decline in the wild. *Science*, **318**, 100–103.

Bernatchez L, Duchesne P (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 1–12.

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, **18**, 503–511.

Bonin A, Bellemain E, Eidesen PB *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Carlin BP, Louis TA (2000) *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, Boca Raton.

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.

Derycke S, Remerie T, Backeljau T *et al.* (2008) Phylogeography of the Rhabditis (Pellioditis) marina species complex: evidence for long-distance dispersal, and for range expansions and restricted gene flow in the northeast Atlantic. *Molecular Ecology*, **17**, 3306–3322.

DeWoody JA (2005) Molecular approaches to the study of parentage, relatedness, and fitness: practical applications for wild animals. *Journal of Wildlife Management*, **69**, 1400–1418.

Dodds KG, Tate ML, McEwan JC, Crawford AM (1996) Exclusion probabilities for pedigree testing farm animals. *Theoretical and Applied Genetics*, **92**, 966–975.

Garber RA, Morris JW (1983) General equations for the average power of exclusion for genetic systems of n codominant alleles in one-parent and in no-parent cases of disputed parentage. In: *Inclusion Probabilities in Parentage Testing* (ed. Walker RH), pp. 277–280. American Association of Blood Banks, Arlington, VA.

Gerber S, Chabrier P, Kremer A (2003) FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes*, **3**, 479–481.

Hixon MA, Pacala SW, Sandin SA (2002) Population regulation: historical context and contemporary challenges of open vs. closed systems. *Ecology*, **83**, 1490–1508.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.

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

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**, 2511–2523.

Jones GP, Milicich MJ, Emslie MJ, Lunow C (1999) Self-recruitment in a coral reef fish population. *Nature*, **402**, 802–804.

Jones GP, Planes S, Thorrold SR (2005) Coral reef fish larvae settle close to home. *Current Biology*, **15**, 1314–1318.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, **16**, 1099–1106.

Kauserud H, Stensrud O, Decock C, Shalchian-Tabrizi K, Schumacher T (2006) Multiple gene genealogies and AFLPs suggest cryptic speciation and long-distance dispersal in the basidiomycete Serpula himantioides (Boletales). *Molecular Ecology*, **15**, 421–431.

Leis JM (2006) Are larvae of demersal fishes plankton or nekton? *Advances in Marine Biology*, Vol 51 **51**, 57–141.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, **20**, 136–142.

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

McLean JE, Seamons TR, Dauer MB, Bentzen P, Quinn TP (2008) Variation in reproductive success and effective number of breeders in a hatchery population of steelhead trout (Oncorhynchus mykiss): examination by microsatellite-based parentage analysis. *Conservation Genetics*, **9**, 295–304.

McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19885–19890.

Meagher TR (1986) Analysis of paternity within a natural population of *Chamaelirium luteum*. 1. Identification of the most-likely male parents. *American Naturalist*, **128**, 199–215.

Milligan BG (2003) Maximum-likelihood estimation of relatedness. *Genetics*, **163**, 1153–1167.

Nathan R (2006) Long-distance dispersal of plants. *Science*, **313**, 786–788.

Nutt KJ (2008) A comparison of techniques for assessing dispersal behaviour in gundis: revealing dispersal patterns in the absence of observed dispersal behaviour. *Molecular Ecology*, **17**, 3541–3556.

Oddou-Muratorio S, Houot ML, Demesure-Musch B, Austerlitz F (2003) Pollen flow in the wildservice tree, Sorbus torminalis

(L.) Crantz. I. Evaluating the paternity analysis procedure in continuous populations. *Molecular Ecology*, **12**, 3427–3439.

Palumbi SR, Grabowsky G, Duda T, Geyer L, Tachino N (1997) Speciation and population genetic structure in tropical Pacific Sea urchins. *Evolution*, **51**, 1506–1517.

Queller DC, Zacchi F, Cervo R *et al.* (2000) Unrelated helpers in a social insect. *Nature*, **405**, 784–787.

Robledo-Arnuncio JJ, Garcia C (2007) Estimation of the seed dispersal kernel from exact identification of source plants. *Molecular Ecology*, **16**, 5098–5109.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Slate J, Marshall T, Pemberton J (2000) A retrospective assessment of the accuracy of the paternity inference program CERVUS. *Molecular Ecology*, **9**, 801–808.

Sokal RR, Rohlf FJ (1995) *Biometry*, 3rd edn. W.H. Freeman and Company, New York.

Team RDC (2007) R: a language and environment for statistical computing. In: *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Telfer S, Piertney SB, Dallas JF *et al.* (2003) Parentage assignment detects frequent and large-scale dispersal in water voles. *Molecular Ecology*, **12**, 1939–1949.

Tentelier C, Guillemaud T, Ferry S, Fauvergue X (2008) Microsatellite-based parentage analysis reveals non-ideal free distribution in a parasitoid population. *Molecular Ecology*, **17**, 2300–2309.

Thompson EA (1975) The estimation of pair-wise relationship. *Annals of Human Genetics*, **39**, 173–188.

Thompson EA (1976) Inference of genealogical structure. *Social Science Information*, **15**, 477–526.

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

Thompson EA, Meagher TR (1998) Genetic linkage in the estimation of pairwise relationship. *Theoretical and Applied Genetics*, **97**, 857–864.

Thorrold SR, Jones GP, Planes S, Hare JA (2006) Transgenerational marking of embryonic otoliths in marine fishes using barium stable isotopes. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 1193–1197.

Vandeputte M, Mauger S, Dupont-Nivet M (2006) An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. *Molecular Ecology Notes*, **6**, 265–267.

Waser PM, Busch JD, McCormick CR, DeWoody JA (2006) Parentage analysis detects cryptic precapture dispersal in a philopatric rodent. *Molecular Ecology*, **15**, 1929–1937.