INVITED REVIEW

Methods of parentage analysis in natural populations

ADAM G. JONES* and WILLIAM R. ARDREN†

*School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332 USA, †Conservation Genetics Laboratory, Abernathy Fish Technology Center, US Fish and Wildlife Service, 1440 Abernathy Creek Road, Longview, WA 98632, USA

Abstract

The recent proliferation of hypervariable molecular markers has ushered in a surge of techniques for the analysis of parentage in natural and experimental populations. Consequently, the potential for meaningful studies of paternity and maternity is at an all-time high. However, the details and implementation of the multifarious techniques often differ in subtle ways that can influence the results of parentage analyses. Now is a good time to reflect on the available techniques and to consider their strengths and weaknesses. Here, we review the leading techniques in parentage analysis, with a particular emphasis on those that have been implemented in readily useable software packages. Our survey leads to some important insights with respect to the utility of the different approaches. This review should serve as a useful guide to anyone who wishes to embark on the study of parentage.

Keywords: exclusion, fractional allocation, likelihood, microsatellites, parentage analysis, paternity

Introduction

An appreciation that molecular techniques can resolve issues of uncertain parentage arose from some of the earliest studies of genetic polymorphisms in populations. Since then, genetic studies of parentage have played a major role in the study of evolution and behavioural ecology and have become one of the central themes in the new field of molecular ecology (Avise 1994; Hughes 1998). Here we focus on the study of parentage in natural populations, which started slowly with the utilization first of chromosomal polymorphisms (Anderson 1974; Milkman & Zeitler 1974; Levine et al. 1980) and later with allozyme electrophoresis (Hanken & Sherman 1981; Ellstrand 1984; Meagher 1986). The advent of DNA fingerprinting in the 1980s led to an explosion of parentage analyses, primarily in birds, that first revealed the power of such studies to overturn existing paradigms in behavioural ecology (Gibbs et al. 1990; Westneat 1990; Birkhead & Möller 1992). At the same time, statistical techniques were being developed for the analysis of parentage using single-locus polymorphisms, such as allozymes (Chakraborty & Hedrick 1983; Meagher & Thompson 1986). The multilocus nature of DNA fingerprinting data precluded simple statistical analysis, so parentage analysis in theory began a noticeable departure from parentage analysis in practice (Pena & Chakraborty 1994). It was the discovery of microsatellite markers that reuni ted theory and practice, resulting in a flood of empirical studies of parentage and a profusion of statistical methods for the analysis of such studies (Luikart & England 1999).

Our goal is to provide a comprehensive guide to the existing methods of analysis, with a particular emphasis on those techniques that have been implemented into readily available computer software packages. Our intent is to direct scientists interested in parentage analysis to the correct set of analytical tools for their particular problems. This review is not intended to address the suitability of various molecular techniques, nor is it meant to review the empirical applications of parentage analysis. Several recent reviews have addressed these issues (Haig 1998; Hughes 1998; Sunnucks 2000; Avise et al. 2002; Wilson & Ferguson 2002). Rather, our intent is to address a topic that has not been reviewed in detail by providing a comparative summary of available methodologies for the analysis of parentage data.

For the purposes of this study, we assume that everyone’s objective is to come as close as possible to complete and perfect parentage assignment. Here we define a complete study of parentage as one in which each and every sampled offspring from a population is assigned its true
mother and father (i.e. the truth is obtained). We recognize that the biological parentage of the offspring per se may not be the goal of most studies of parentage. Rather, reconstruction of parentage is usually a means to achieve the goal of evaluating some specific ecological hypothesis and a perfect parentage analysis may not be necessary to evaluate many hypotheses. However, existing techniques of parentage reconstruction share the characteristic that they all attempt (at some level) to reconstruct patterns of parentage in the population. Thus, they are compared most easily if we consider how they differ in their methodologies for achieving this goal.

**Methods of parentage analysis**

**Exclusion**

The earliest and conceptually simplest technique of parentage analysis is exclusion. This technique (based on Mendelian rules of inheritance) uses incompatibilities between parents and offspring to reject particular parent-offspring hypotheses (see Information box). Exclusion is an appealing approach, because exclusion of all but one parent pair from a complete sample of all possible parents for each offspring in a population could be considered the paragon of parentage analysis. However, few studies have achieved this ideal. One of the potential weaknesses of a strict exclusion approach is that genotyping errors, null alleles and mutations will contribute to false exclusions (see below). Ironically, these problems become more acute as more data are brought to bear on a problem, because the assay of additional loci (or additional individuals) increases the likelihood that a dataset will contain errors or mutations.

**Categorical and fractional likelihood**

If complete exclusion is not possible, then the researcher must resort to other methodologies. Hence, techniques were developed that assigned progeny to nonexcluded parents based on likelihood scores derived from their genotypes (see Information box). These techniques assign offspring either categorically or fractionally. Both categorical and fractional allocation techniques calculate the likelihoods in essentially the same way. They differ only in that the categorical technique assigns the entire offspring to a particular male, whereas the fractional technique splits an offspring among all compatible males. Intuitively, the categorical technique appears to have the upper hand (and indeed has been a more popular approach), because it has the potential to produce results that represent biological truth. In contrast, the fractional technique is guaranteed to be incorrect from a biological standpoint, because an offspring can have only one father and one mother — there is no such thing as fractional parentage. Despite the intuitive appeal of the categorical technique, however, fractional assignment possesses better statistical properties for the evaluation of some hypotheses than the categorical assignment technique. Consequently, even though the fractional technique may not represent the exact biological truth, it can provide more exact estimates of important mating system parameters from a statistical standpoint. For example, fractional techniques appear to have great promise for less biased estimates of the proportion of offspring in a population parented by each of the adults (Devlin et al. 1988; Smouse & Meagher 1994; Neff et al. 2001), for comparing the reproductive success of different categories of males (Nielsen et al. 2001; Signorovitch & Nielsen 2002) and for incorporating prior information about the biology of the species into the analysis (Neff et al. 2001; Nielsen et al. 2001).

One interesting point regarding likelihood-based parentage techniques comes from inspection of the likelihood equations (see Information box). Not everyone appreciates that the only information regarding the likelihood of parentage for multiple males that are compatible with a single offspring comes from whether the males are homozygotes or heterozygotes. The homozygous males typically share more alleles with a putative offspring than the heterozygous males, because homozygotes have two copies of the compatible allele. Consequently, a male’s likelihood of paternity is positively related to his homozygosity. It is important to note that this positive relationship is a statistical reality and should not be viewed as a weakness of the technique — a homozygous male that is compatible with a given offspring really does have a higher likelihood of producing it than a heterozygous male sharing one allele with the offspring. Nevertheless, not much information on the relative likelihoods of two compatible males can be gleaned from a small number of loci, making it imperative that a technique be used to assess statistical confidence in parentage assignments (see below).

The importance of understanding the assumptions underlying the categorical and fractional techniques is best illustrated by an example. One common goal of parentage analyses is to estimate the variances in reproductive or mating success for one or both sexes in a population. The fractional and categorical techniques are expected to perform differently from one another in this enterprise. The fractional technique requires the researcher to set a prior probability of parentage (see Information box). Because little data are typically available to determine these prior probabilities, most researchers chose a uniform distribution, which specifies that all males have equal probabilities of paternity (Devlin et al. 1988). However, this uniform prior will almost always result in an underestimate of variance in reproductive success (Neff et al. 2001). On the other hand, the tendency of the categorical method to overestimate the reproductive success of individuals with
Information Box: approaches for calculating parentage

Exclusion

The process of exclusion (based on Mendelian rules of inheritance) uses incompatibilities between parents and offspring to reject particular parent–offspring hypotheses. For example, if a mother and offspring have the diploid genotypes A/A and A/B, respectively, at a single locus, then males with the genotype A/C can be excluded whereas those with the genotype B/C cannot. This technique is most powerful when there are few candidate parents and highly polymorphic genetic markers available. However, this method can become impractical when the pool of candidate parents becomes large due to the large number of loci needed to yield a single nonexcluded parent or parent-pair assignment to an offspring. Under strict exclusion, a single mismatch is enough to exclude a candidate parent. However, many exclusion programs can allow the user to specify the number of mismatches necessary for an exclusion to be considered valid, making the method more robust to the difficulties imposed by mutations or scoring errors.

Categorical allocation

Categorical allocation uses likelihood-based approaches to select the most likely parent from a pool of nonexcluded parents. This method involves calculating a logarithm of the likelihood ratio (LOD score) by determining the likelihood of an individual (or pair of individuals) being the parent (or parents) of a given offspring divided by the likelihood of these individuals being unrelated. After an exhaustive evaluation of all genetically possible parents, offspring are assigned to the parent (or parental pair) with the highest LOD score. When all parent–offspring relationships show zero likelihood, offspring are unassigned. Parentage remains ambiguous when multiple parent–offspring relationships obtain the highest nonzero likelihood. Contrary to strict exclusion methods, likelihood-based allocation methods usually allow for some degree of transmission errors due to genotype misreading or mutation.

We present Meagher & Thompson’s (1986) original likelihood-based approaches for determining the most probable single parents and parent pairs. For consistency, we use the notation of Meagher & Thompson (1986). In all cases, we examine genotypes ($g_B$, $g_C$, and $g_D$) at a single autosomal locus for three individuals (B, C and D). Assuming loci are unlinked, information from multiple loci can be combined by summing the LOD scores over all loci. Variations or extensions of these original approaches have been developed and implemented (Sancristobal & Chevalet 1997; Marshall et al. 1998; Gerber et al. 2000). Transition probabilities (T) for use in the following equations can be found in Marshall et al. (1998) for codominant markers and in Gerber et al. (2000) for dominant markers. Our headings correspond to the descriptions of these approaches used by Sancristobal & Chevalet (1997).

(a) Identifying one parent when the other is known. Letting C represent the known parent and D the alleged parent, the LOD score for D being the parent of B is:

\[
\text{LOD score (D parent of B)} = \log_{10} \frac{T(g_B | g_C, g_D)}{T(g_B | g_C)}
\]  

where $T(g_B | g_C, g_D)$ is the transition probability of $g_B$ given $g_C$ and $g_D$ and $T(g_B | g_C)$ is the transition probability of $g_B$ given $g_C$.

(b) Identifying one parent with no information about the other parent. In this case, no information is available concerning parentage of B. The single parent LOD score for C being the parent of B is:

\[
\text{LOD score (C parent of B)} = \log_{10} \frac{T(g_B | g_C)}{P(g_B)}
\]  

where $P(g_B)$ is the frequency of the offspring’s genotype in the population.

(c) Identifying a parental pair starting with no prior information. Parental pair allocation is an approach for identifying parent–offspring relationships by constructing genotypic triplets consisting of a proposed offspring and proposed maternal and paternal parents. This procedure involves calculating a breeding likelihood, which is defined as the likelihood of a parental pair producing the multilocus genotype found in the offspring being examined. The breeding likelihood of a given offspring on the basis of a single locus is:

\[
\text{LOD score (C, D parents of B)} = \log_{10} \frac{T(g_B | g_C, g_D)}{P(g_B)}
\]

Fractional allocation

The fractional allocation method assigns some fraction, between 0 and 1, of each offspring to all nonexcluded candidate parents. The portion of an offspring allocated to a particular candidate parent is proportional to its likelihood of parenting the offspring compared to all other nonexcluded candidate parents. Single parent and parent pair likelihoods are calculated in the same way as in the categorical allocation method. We present
the original fractional paternity approach suggested by Devlin et al. (1988). This approach assumes genotypes are known from all parents in the population and that one parent is known for the offspring under consideration. The fraction of offspring (O = k) awarded to a particular candidate male j (MP = j) on female parent i (FP = i), is denoted by Fij and is estimated by:
\[
F_{ij} = \sum_k X_{ik} P(\text{MP} = j | \text{FP} = i, O = k)
\] (4)
where \(X_{ik}\) is the number of offspring of each distinct genotype from this female parent. Bayes’ Theorem can be used to estimate the probability that the candidate male parent (\(j^*\)) is the actual male parent, given the female parent and genotype of the offspring (Devlin et al. 1988):
\[
P(\text{MP} = j^* | \text{FP} = i, O = k) = \frac{T(\gamma_k | \alpha_j, \beta_p) P(\text{MP} = j^* | \text{FP} = i)}{\sum_j T(\gamma_k | \alpha_j, \beta_p) P(\text{MP} = j | \text{FP} = i)}
\] (5)

The vectors \(\alpha_i, \beta_j, \beta_p\) and \(\gamma_k\) represent the multilocus genotypes for the female parent, candidate male j from the population, the putative male under consideration, and the offspring, respectively. The prior probability of paternity, \(P(\text{MP} = j | \text{FP} = i)\) takes into account all of the behavioural, ecological and genetic parameters determining the likelihood of each candidate male being the parent of an offspring from female i. Neff et al. (2001) provide a useful summary of the types of functions that can be used to estimate \(P(\text{MP} = j | \text{FP} = i)\) and other prior probability parentage vectors. If it is assumed \(P(\text{MP} = j | \text{FP} = i)\) is constant (i.e. each candidate parent is equally likely to fertilize a given offspring) for all j then equation 5 is simply:
\[
P(\text{MP} = j^* | \text{FP} = i, O = k) = \frac{T(\gamma_k | \alpha_j, \beta_p)}{\sum_j T(\gamma_k | \alpha_j, \beta_p)}
\] (6)

Devlin et al. (1988) used eqn 6 as the basis for fractionally allocating offspring among all of the candidate male parents. This model is easily modified to situations where neither parent is known.

Many variations on the fractional allocation method have been developed. Roeder et al. (1989) used genetic likelihood-based procedures to estimate individual fertilities, and Smouse & Meagher (1994) developed methods to estimate the distribution of male reproductive success within a population. Nielsen et al. (2001) presented a method that accommodates incomplete sampling of candidate parents and evaluates relative reproductive success of groups within the population. Another method developed by Neff et al. (2001) incorporates other biological data, such as behavioural observations, to estimate the prior probability distribution of parentage and allows for confidence intervals to be estimated for the parentage assignments.

**Parental reconstruction**

This method uses the multilocus genotypes of parents and offspring to reconstruct the genotypes of unknown parents contributing gametes to a progeny array for which one parent is known a priori (Jones 2001). Associations of alleles across loci provide information regarding the genotypes of parents contributing to a progeny array (Jones & Avise 1997). Existing techniques reconstruct the minimum number of parental genotypes necessary to explain the data set, using an exhaustive algorithm (Jones 2001). For the case in which the mother is known, all possible paternal genotypes consistent with at least one progeny in the data set are tested in combination to determine which minimum set of paternal genotypes can explain the entire progeny array. This technique is extremely computationally intensive, and becomes prohibitively time-consuming for progeny arrays with more than about six fathers.

many homozygous loci and to underestimate the reproductive success of individuals with many heterozygous loci causes an upward bias in the estimate of variance in reproductive success (Devlin et al. 1988; Smouse & Meagher 1994). Clearly, results from parentage studies using perfect exclusion would provide the most robust estimates of variance in reproductive success, but such studies can be difficult to perform. Another solution is to use a fractional assignment technique with the appropriate prior probability of parentage (Neff et al. 2001), but (as noted above) these priors are rarely known with certainty in natural systems. Regardless, this example illustrates that the results of parentage analyses must be carefully considered in terms of the strengths and weaknesses of the analytical techniques.

**Genotypic reconstruction**

A final approach to parentage analysis is to reconstruct parental genotypes from progeny arrays of full- or half-sibs (see Information box). These reconstructed genotypes can then be compared to the genotypes of a pool of

© 2003 Blackwell Publishing Ltd, Molecular Ecology, 12, 2511–2523
candidate parents or can be compared to one another to detect multiple mating even with a complete absence of genetic samples from one sex (Jones & Avise 1997; Jones et al. 1998a). As with the other techniques, this one is susceptible to scoring errors, null alleles and mutations. However, research involving progeny arrays has an enormous advantage in this respect, because these sources of complication often will affect only one or a few offspring at a single locus. Consequently, the unexpected genotypes of the affected offspring will usually raise eyebrows, allowing for appropriate diagnosis of the source of the inconsistencies.

**Choosing the appropriate parentage analysis method**

**Collecting constraints**

The appropriate technique for data analysis will be dictated in large part by the types of samples that can be obtained from the study system (Table 1). The best-case scenario would be one in which large groups of offspring could be collected from known mated pairs of adults. In such a situation, molecular techniques need be employed only to verify that the suspected relationships are true (Gowaty & Karlin 1984; Gyllensten et al. 1990). The situation becomes more difficult as the completeness of the sample diminishes (Table 1). For example, when offspring can be collected in family groups with their mothers, and a complete sample of adult males from the population can be obtained, the prospects for complete assignment of parentage are excellent. As certain favourable features of this sample are lost, the likelihood of a successful study drops. If not all candidate parents can be sampled, for instance, correct assignment of parentage will be impossible for some of the progeny, and some progeny may be assigned incorrectly. In addition, if offspring cannot be collected in family groups, one entire analytical approach (i.e. parental reconstruction) will not be possible at all. As the sample deviates further and further from the ideal, parentage analysis in the strict sense becomes unachievable. Under such circumstances, the data may still be useful for the estimation of some mating system parameters, such as the prevalence of multiple paternity or the rate of selfing in plants (Ellstrand 1984; DeWoody et al. 2000; Ritland 2002). It is also important to realize that as the sample becomes less favourable for parentage analysis (i.e. as one moves down the rows in Table 1), greater resolving power of the molecular marker system will be required for a successful study.

One sampling constraint, the proportion of candidate parents that can be sampled, deserves special attention. A partial sample of candidate parents produces two major types of problems that cannot be overcome easily with existing techniques. First, assignment techniques require knowledge of the total number of candidate parents in the population, because this value plays a major role in the assessment of confidence in assignments (Marshall et al. 1998; Nielsen et al. 2001). The results of programs such as CERVUS and FAMOZ are extremely sensitive to the estimate of the proportion of candidate adults sampled (which is provided as a parameter by the user). The total number of breeding adults in a population is rarely known, so this problem is a major concern for many types of studies. In an effort to alleviate this problem, PATRI can use genetic data to estimate the total number of breeding adults in the population and can incorporate uncertainty regarding population size into the analysis (Signorovitch & Nielsen 2002). The second problem is specific to programs that assign parent pairs, such as PAPA or PROBMAX. The difficulties associated with incomplete sampling of parents are exacerbated by the use of a parent-pair assignment algorithm, because a failure to sample either member of a breeding pair will render correct assignment impossible. Thus, if 50% of adults can be sampled, we might expect only about 25% of the breeding pairs to appear in the sampled adults. Thus, a parent-pair technique will be able to assign fathers for at most about 25% of the offspring. This requirement to assign both parents simultaneously also makes the parent-pair technique especially vulnerable to null alleles, mutations and scoring errors.

**Molecular marker considerations**

Once appropriate tissue samples from the focal population have been obtained, how should they be analysed genetically? For most biological systems, the most powerful genetic tools for parentage analysis will be microsatellite markers, and most of the recent advances in techniques of data analysis have been aimed at studies employing microsatellites. Once a battery of molecular markers has been developed for a particular system, exclusion probabilities can be calculated and used as a rough guide to which and how many loci should be used for the actual analysis. Exclusion probabilities have been defined and discussed at length elsewhere (Chakraborty et al. 1988; Dodds et al. 1996), so we will not belabour them here. However, many of the parentage analysis programs include subroutines that calculate exclusion probabilities (Table 2).

For some systems, markers other than microsatellites may be more tractable and can be employed profitably. This review is written from the perspective of a scientist using microsatellite markers, but the analytical techniques can be applied to any codominant marker. In addition, we point out accommodations that programs have made for other types of markers, such as dominant amplified fragment length polymorphisms (AFLPs) or uniparentally inherited cytoplasmic markers (Gerber et al. 2000). Once
Table 1 The implications of various types of samples from natural populations for the reconstruction of parentage

<table>
<thead>
<tr>
<th>How many of the unknown parents can I sample?</th>
<th>Can I sample large groups of full or half-sibs?</th>
<th>How many parents of each offspring are known a priori?</th>
<th>Preferred technique</th>
<th>What can I do if I don’t powerful have sufficiently markers for the preferred technique?</th>
<th>Possible computer programs (see Table 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All or some</td>
<td>Yes or no</td>
<td>Both</td>
<td>Genotype some individuals to verify parent-offspring relationships</td>
<td>Nothing</td>
<td>Newpat, cervus, famoz, parente, patri of kinship for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>All</td>
<td>Yes</td>
<td>One</td>
<td>Exclusion corroborated by reconstruction of parental genotypes from progeny arrays</td>
<td>Categorical allocation, fractional allocation, estimate number of parents</td>
<td>Probmax, papa, famoz or parente for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>All</td>
<td>Yes</td>
<td>Neither</td>
<td>Complete exclusion. If progeny arrays contain half-sibs, reconstruction of parental genotypes</td>
<td>Categorical allocation, fractional allocation, estimate number of parents</td>
<td>Newpat, cervus, famoz, parente, patri of kinship for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>All</td>
<td>No</td>
<td>One</td>
<td>Complete exclusion</td>
<td>Categorical allocation, fractional allocation, kinship techniques</td>
<td>Newpat, cervus, famoz, parente, patri of kinship for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>All</td>
<td>No</td>
<td>Neither</td>
<td>Complete exclusion</td>
<td>Categorical allocation, fractional allocation, kinship techniques</td>
<td>Newpat, cervus, famoz, parente, patri of kinship for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>Some</td>
<td>Yes</td>
<td>One</td>
<td>Reconstruction of parental genotypes. Complete exclusion or categorical/fractional allocation</td>
<td>Categorical allocation, fractional allocation, estimate number of parents</td>
<td>Newpat, cervus, famoz, parente, kinship, probmax1 or papa1 for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>Some</td>
<td>Yes</td>
<td>Neither</td>
<td>Complete exclusion or categorical/fractional allocation. Reconstruction of parental genotypes if progeny arrays contain half-sibs</td>
<td>Categorical allocation, fractional allocation, estimate number of parents</td>
<td>Newpat, cervus, famoz, parente, kinship, probmax1 or papa1 for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>Some</td>
<td>No</td>
<td>One or neither</td>
<td>Complete exclusion or categorical/fractional allocation.</td>
<td>Kinship techniques</td>
<td>Newpat, cervus, famoz, parente, patri (if one parent is known), kinship, probmax1 or papa1 for exclusion or allocation. Gerud for reconstruction</td>
</tr>
<tr>
<td>None</td>
<td>Yes</td>
<td>One</td>
<td>Reconstruction of parental genotypes</td>
<td>Estimate number of parents</td>
<td>Gerud for reconstruction</td>
</tr>
<tr>
<td>None</td>
<td>Yes</td>
<td>Neither</td>
<td>Reconstruction of parental genotypes if progeny arrays contain half-sibs</td>
<td>Estimate number of parents</td>
<td>Gerud for reconstruction</td>
</tr>
<tr>
<td>None</td>
<td>No</td>
<td>One or neither</td>
<td>Use kinship or relatedness techniques</td>
<td>Nothing</td>
<td>Kinship of Relatedness2</td>
</tr>
</tbody>
</table>

1Parent-pair allocation techniques are particularly sensitive to incomplete sampling of candidate parents and should be used carefully here.

2RELATEDNESS is available from http://gsoft.smu.edu/
<table>
<thead>
<tr>
<th>Method</th>
<th>Program</th>
<th>PM</th>
<th>PP</th>
<th>PR</th>
<th>IC</th>
<th>EC</th>
<th>EP</th>
<th>Required info</th>
<th>Type of genetic markers</th>
<th>Ability to accommodate ‘error’</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusion</td>
<td>probmax5</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of offspring and sexed parents</td>
<td>Diploid codominant</td>
<td>Good</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>newpat4</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of offspring and sexed parents</td>
<td>Diploid dominant</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>kinshipa</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of offspring and sexed parents</td>
<td>Diploid codominant</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>cervusi2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of offspring and sexed parents</td>
<td>Diploid codominant</td>
<td>Moderate</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>papa1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of offspring and parents</td>
<td>Diploid codominant</td>
<td>Poor</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>famoz2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>Genotypes of offspring and parents</td>
<td>Diploid codominant</td>
<td>Poor</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>parente3</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of offspring and parents</td>
<td>Cytoplasmic</td>
<td>Poor</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>patri5</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of parent-offspring and sexed parents</td>
<td>Diploid codominant</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>gerudα</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Genotypes of known parent with a large group of its progeny</td>
<td>Diploid codominant</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Purity/maternity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Can test relative reproductive success of different groups</td>
</tr>
<tr>
<td>Parental pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Can also be used for categorical allocation</td>
</tr>
<tr>
<td>Reconstruction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Uses multilocus data to determine the minimum number of sires for a family</td>
</tr>
</tbody>
</table>

*a*Paternity/maternity (see Information box).
*b*Parental pair allocation (see Information box).
*c*Parental reconstruction (see Information box).
*d*Ability to assign statistical confidence for particular parent–offspring pairs.
*e*Ability to assess the expected confidence in assignments on an experiment-wide basis.
*f*Ability to calculate exclusion probabilities.

5*Danzmann (1997); available at http://www.uoguelph.ca/~rdanzman/software/probmax/*
6*Worthington Wilmer et al. (1999); available at http://www.zoo.cam.ac.uk/zoostaff/amos/newpat.html*
7*Goodnight & Queller (1999); available at http://gsoft.smu.edu/*
8*Marshall et al. (1998); available at http://helios.bto.ed.ac.uk/evolgen/cervus/cervus.html*
9*Duchesne et al. (2002); available at http://www.bio.ulaval.ca/contenu-fra/professeurs/Prof-l-bernatchez.html*
10*Gerber et al. (2000); available at http://www.pierroton.inra.fr/genetics/lab0/software/Famoz/*
11*Cercuell et al. (2002); available at http://www2.ujf-grenoble.fr/leca/membres/manel.html*
12*Signorovitch & Nielsen (2002); available at http://www.biom.cornell.edu/Homepages/Rasmus_Nielsen/files.html*
13*Jones (2001); available at http://www.biology.gatech.edu/professors/labsites/jones/parentage.html*
the genetic data have been collected, they will be analysed typically by a computer program, except in the simplest of experiments.

**Choice of computer software**

With respect to the statistical analysis of the data, the choice of technique is again governed primarily by constraints imposed by the sample. In Table 1 we suggest which programs are likely to be useful for the analysis of each type of sample that can be obtained from a natural population. One point to keep in mind is that likelihood and fractional assignment techniques should be seen as ways of compensating for shortcomings of the data set, which can arise as a result of insufficient genetic variation, scoring errors, mutations, null alleles or incomplete sampling. As a consequence, those programs that perform likelihood or fractional analyses can also usually be used for brute-force exclusion. The majority of computer programs use some sort of likelihood algorithm to assign parentage categorically (Table 2). However, other potentially useful programs focus on exclusion, fractional assignment techniques or parental genotypic reconstruction (Table 2). Parentage analysis software packages differ importantly with respect to the types of problems that they were designed to analyse and we summarize these differences in Table 2. Perhaps the most important differences among programs stem from how they handle sources of difficulty in parentage analysis, such as null alleles, mutations, and scoring errors.

**Technical and biological hurdles for parentage analysis**

**Null alleles**

Microsatellite null alleles result typically from polymorphism in the flanking sequence of the locus, such that some alleles lack a functional polymerase chain reaction (PCR) priming site (Callen *et al*. 1993; Jones *et al*. 1998b). Null alleles are an important consideration for parentage analysis, because they can cause false exclusions when null heterozygotes are scored incorrectly as non-null heterozygotes. For example, if an offspring has the genotype A/null, it will be scored as A/A and will be deemed incompatible with B/null and C/null fathers (scored B/B and C/C), even though in reality it is compatible with these males. Fortunately, null alleles can usually be detected as a significant departure from Hardy–Weinberg equilibrium. In studies in which a known parent is sampled with groups of offspring, null alleles are even easier to detect because they result in incompatibilities between the known parent and offspring that invariably involve homozygous genotypes.

Several of the computer programs are capable of detecting null alleles, but few deal effectively with them in the analysis (Table 2). The most conservative way to handle a locus with null alleles in parentage analysis is to recode all homozygous genotypes as heterozygotes possessing the detected allele and the null allele, thus preventing exclusion on the basis of homozygous genotypes. PROBMX treats loci known to have null alleles in this way. On the other hand, CERVUS, FAMOZ, PAPA and PARENTE deal with null alleles by treating them as any other mutation or genotyping error (see below). This approach is problematic, because these programs allow only a single, experiment-wide mistake rate that is constant across loci. Thus, if one locus has a high frequency null allele, then the program can accommodate it only by assuming that all loci have a high rate of mistakes or mutations, thus diminishing the power of the loci that do not display null alleles. Consequently, all these programs handle null alleles poorly in their analyses. CERVUS, however, does run an algorithm that is capable of detecting deviations from Hardy–Weinberg equilibrium, and it warns the user when null alleles are likely to be present. FAMOZ and PAPA possess no such algorithm. NEWPAT uses a likelihood ratio test to attempt to identify offspring with null alleles. This approach does not perform as well as the approach used by PROBMX, although it may ameliorate somewhat the effects of null alleles. For now, the best solution for exclusion and assignment analyses is to recode all genotypes (manually if necessary) in a way similar to that used by PROBMX for any locus suspected of having a null allele. For the reconstruction of parental genotypes it is currently better to avoid using loci with null alleles.

**Linked loci and linkage disequilibrium**

Most parentage analysis techniques consider the data one locus at a time and combine the information over all loci by assuming independent assortment. The linkage relationships of multiple microsatellite loci are characterized rarely, due to limited resources or lack of appropriate samples. For studies employing a small number of loci, the probability that they will be physically linked or in linkage disequilibrium (i.e. gametic phase disequilibrium) is usually small, so most researchers assume independent assortment of alleles among loci. In fact, none of the existing computer programs for parentage analysis make accommodations for linked loci or linkage disequilibrium. However, physically linked loci have been encountered in several studies of parentage in nature (Jones *et al*. 1998c, 2001; Ardren *et al*. 1999), demonstrating that the assumption of no linkage among loci does not always hold.

The effects of linkage disequilibrium and physical linkage between loci are different in regard to parentage analysis. Using loci that are in linkage disequilibrium decreases
Mutations and scoring errors

Mutations and genotyping errors are additional complications for studies of parentage. Most programs handle them as a single class of error, so we deal with them simultaneously here. Mutations and errors are obviously of potential importance for parentage studies, because they can cause offspring to appear incompatible with their true biological parents. Not all parentage algorithms make allowances for mutations and mistakes (Table 2). However, these programs are still useful, because a mistake- and mutation-free data set is not out of the realm of possibility. Mutation rates at microsatellite loci usually are low enough that a data set including several thousand genotypes will contain few mutations (Ellegren 2000). The data sets most conducive to avoiding errors are those in which large groups of progeny can be collected with a known parent, because inconsistencies between parents and offspring as well as among siblings will usually raise a red flag whenever a mutation or error occurs. Data sets in which offspring and parents are collected singly and separately from one another within a population are most vulnerable to undetected mutations, because no check of the transmission of alleles between relatives is possible in such a sample.

For those data sets in which mutations and scoring errors appear to be an important concern, algorithms have been developed to overcome such problems (Sancristobal & Chevalet 1997; Marshall et al. 1998). The details of these algorithms differ among programs. For example, in CERVUS mistakes are assumed to replace an entire single-locus genotype with another genotype according to its expected frequency in the population. Thus, an assignment is more likely to be allowed if a mismatch between a putative father and an offspring involves a common genotype than if the mismatch involves a rare genotype. PARENTE handles mismatches in essentially the same way as CERVUS, except that the single allele causing the mismatch is replaced with an allele that is chosen according to its frequency in the population. PAPA uses a distinct algorithm developed by Sancristobal & Chevalet (1997) that allows mistakes and mutations to be drawn from a specified distribution that can range from a strict stepwise model to an infinite allele model, depending on parameter values specified by the user. One difficulty with all such attempts to accommodate mutations is that in very few cases will sufficient data be available to allow an appropriate choice of a model of mutation. Other programs use slightly simpler algorithms. For example, FAMOZ uses the method of Sancristobal & Chevalet (1997), but restricts its use to the infinite alleles model. The exclusion program NEWPAT simply allows a user to specify the number of mismatches that are necessary for exclusion. PARENTE also allows the user to specify the maximum allowable number of mismatches for an assignment, but it currently is not clear how the interaction between this parameter and PARENTE’s error-handling technique will affect the results of parentage assignment. PROBMAX can rank putative parents by their degree of compatibility with the offspring in question and it can also allow small mistakes within a stepwise mutation framework. KINSHIP, PATRI and GERUD do not make any allowances for genotyping errors or mutations, but do detect incompatibilities between known parents and offspring, which can give some indication of the rate of typing error in the data set. Given our current limited knowledge of the nature of mistakes and mutations at microsatellite loci, it is difficult to predict which error-handling algorithm will give the most reliable results in actual analyses.

Extended family structure in the population

The presence of family members other than the parents of the offspring in the pool of candidate parents can present a serious challenge to parentage assessment. Most empirical studies of parentage assume that sampled adults are not related to one another and that no relatives of the offspring other than parents are included in the sample of adults. Violation of this assumption can have a major effect on the prospects for a successful parentage analysis. The most problematic situation arises when either half- or full-siblings of some of the progeny are included in the pool of candidate parents. A slightly less severe problem arises when some of the candidate parents are related to one another. Even though some analyses of the effects of family structure on parentage assignment have been performed (Thompson & Meagher 1987; Marshall et al. 1998; Nielsen et al. 2001), additional work will be necessary before we
can say with any confidence which techniques are most susceptible to violations of assumptions about family structure. One certainty, however, is that a strict exclusion approach will still be valid even in the face of extremely high levels of relatedness. The problem is that if relatedness values are too high, the number of loci required for exclusion could be prohibitive. It is clear that we need additional research on the extent to which important family structure is expected to prevail in natural populations and on the sensitivity of different parentage analysis techniques to extended family structure.

Assessing statistical confidence in parentage assignments

Most systems in nature do not permit a perfect parentage analysis through complete exclusion. If it is necessary to resort to other statistical methods of parentage analysis, then a major goal should be to assess to what degree a researcher should feel confident about the reliability of parentage assignments. Thus, one extremely important contribution has been the simulation-based assessment of confidence in assignments used by \textit{cervus}. For each analysis, \textit{cervus} simulates data sets and calculates expected distributions of their test statistic, $\Delta$, which is usually the difference in likelihood ratios between the two males most likely to be the father of the offspring. From the distributions of the test statistic in simulations, \textit{cervus} can determine a critical value that will produce a desired level of confidence in assignments of parentage. This algorithm is particularly important when mutations and mistakes exist in the data set, because under such circumstances all hypothesized parent–offspring pairings possess a finite probability (sometimes vanishingly small) of being true, and \textit{cervus} provides an objective means by which the validity of pairings possessing some apparent genetic incompatibilities can be evaluated. \textit{famoz} uses the same algorithm as \textit{cervus} for establishing confidence in assignments, and it also adds some additional features, such as the ability to use dominant and cytoplasmic markers and to assign parent pairs simultaneously. \textit{papa} and \textit{cerud} also use simulations to assess expected success in parentage assignment (or reconstruction) on an experiment-wide basis, but are more limited than \textit{cervus} and \textit{famoz} because they do not provide confidence scores for particular assignments of parentage at the level of individuals.

\textit{KINSHIP} uses a similar method to that employed by \textit{cervus} and \textit{famoz} for determining confidence in the results of parentage assignment. \textit{KINSHIP} was designed to test a wide range of hypothesized pairwise relationships among individuals, and parentage analysis is one of its many possible applications. As in \textit{cervus}, \textit{KINSHIP} calculates a likelihood ratio of the hypothesized relationship (i.e. parent–offspring) to the null hypothesis (i.e. unrelated) for each candidate parent–offspring dyad. The program then performs a simulation to determine a critical value for this likelihood ratio that produces a desired level of confidence in the results. One major difference between this approach and the one used by \textit{cervus} is that \textit{cervus} uses the difference in the log of the likelihood ratios of the most likely male and the second most likely male as its test statistic, whereas \textit{KINSHIP} simply uses the raw likelihood ratio. In addition, the approach used by \textit{cervus} simulates simultaneously an entire population of candidate males equal in size to the focal population under study and asks what critical value of $\Delta$ is necessary for confident assignment. \textit{KINSHIP}, on the other hand, simulates individual parent–offspring pairs one by one. In essence, \textit{cervus} corrects for multiple comparisons in the study and \textit{KINSHIP} does not. Consequently, the levels of confidence reported by \textit{KINSHIP} will tend to be much higher than those reported by \textit{cervus}. Additional research will be necessary to assess the relative performance of \textit{KINSHIP} vs. other techniques, but the assessment of statistical confidence provided by \textit{KINSHIP}, because it does not correct for multiple comparisons, should be interpreted cautiously if this program is applied to the study of parentage.

The simulation-based approach to confidence assessment implemented by \textit{cervus} and \textit{famoz} has been criticized by Nielsen \textit{et al}. (2001). The major criticism is that the test statistic $\Delta$ only uses information from the two most likely males, discarding the information from other compatible males. In addition, the value of $\Delta$ is calculated differently when only one male is compatible with the offspring than when two or more are compatible, presumably resulting in two different distributions of $\Delta$, but \textit{cervus} does not treat these cases separately in calculating critical values. In answer to the weaknesses of the simulation-based approach, \textit{patri} uses a Bayesian method, in which the posterior probability of paternity is calculated for each father using information from all possible fathers (sampled and unsampled) in the population. The equation for this posterior probability is similar to the equation for the fractional paternity likelihood (Information box), but with an additional term added to account for incomplete sampling of candidate parents (Nielsen \textit{et al}. 2001). For each mother–offspring pair, \textit{patri} produces a posterior probability of paternity for each compatible male. \textit{parente} uses essentially the same Bayesian posterior probability as \textit{patri}, but with accommodations for errors and parent-pair parentage assignment. For both these programs, the posterior probability serves as a guide to which parents are most likely. However, the interpretation of these posterior probabilities is not entirely clear, because neither of these programs provides critical values for the probabilities that will produce a desired level of experiment-wide error. Additional research will be necessary to establish functional criteria for the use of posterior probabilities.

\textit{newpat} takes an entirely different approach to paternity assignment. It reports two values that are intended to serve as
a guide to which of the nonexcluded sires is most likely. First, NEWPAT calculates the relatedness (Queller & Goodnight 1989) between the potential father and the offspring. Second, the program creates a file of random ‘pseudomales’ (based on allele frequencies) and asks what percentage of these pseudomales would not be excluded for each offspring in the data set. The value of these methods for assessing reliability in assignments is not clear. Worthington Wilmer et al. (1999) provide no clear empirical or theoretical justification for this approach. In addition, the program does not provide any type of critical value or any means for combining the values obtained by the relatedness and randomization techniques for assessing confidence. The bottom line is that this technique should not be used for determining confidence in assignments until additional research has been conducted to test its validity. NEWPAT uses a less questionable approach to assess experiment-wide error in which it simulates data sets and calculates the background level of paternities expected by chance (assuming the males are unrelated to the offspring).

**Statistical evaluation of hypotheses without explicit parentage assignment**

While we have focused mainly on ways of reconstructing patterns of parentage as if it is the ultimate goal of a study, parentage analysis is usually motivated by a desire to calculate some population-level parameter, such as the intensity of sexual selection or the frequency of extra-pair fertilizations. A promising new approach is to estimate population parameters of interest directly from the genotypic data. This goal seems to be the most appropriate application of the fractional assignment methods. For example, PATRI uses the fractional technique to evaluate various hypotheses regarding the reproductive success of males from different groups. Another recent promising application of the fractional method permits the estimation of mating system parameters in a way that provides confidence intervals and incorporates prior information about the population (Neff et al. 2001), such as observations of dominance hierarchies or mating interactions. Other techniques have been developed to estimate directly gene flow (Devlin & Ellstrand 1990; Adams et al. 1992), selection gradients (Smouse et al. 1999; Morgan & Conner 2001) and reproductive success of nest-holding males (Neff et al. 2000). Further refinement of these techniques and the development of software to facilitate their widespread use will no doubt be extremely influential in future studies of parentage in natural populations.

**Conclusions**

The current outlook in the field of parentage analysis is extremely positive. Sufficiently powerful genetic markers can be developed in most systems for very complete parentage analysis. The analytical tools at our disposal are also excellent, allowing several different approaches to parentage reconstruction. One of the most important recent advances is the merging of likelihood techniques with techniques that assess confidence in assignments. Some remaining drawbacks with existing computer packages are that they generally do a poor job of handling null alleles and their success at handling mutations and errors is not well understood. In addition, there may be room for improvement in the assessment of statistical confidence in parentage assignments. Nevertheless, it appears that the major remaining challenge in parentage analysis is to obtain appropriate and complete field samples, so one encouraging point (for those among us who are field-orientated) is that more effort should be devoted to creative sampling practices (i.e. spend more time in the field).

**Acknowledgements**

We would like to thank M. Blouin for spurring us to write this review. We are also grateful to three anonymous referees, who provided valuable comments on the manuscript.

**References**


Danzmann RG (1997) PROBMAX: a computer program for assigning unknown parentage in pedigree analysis from known


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


Research in the Jones lab addresses basic questions in ecology and evolution through the use of molecular and computational techniques. W. Ardren maintains an active research program focused on the population genetics and conservation of anadromous and freshwater fish.