MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 6–30 doi: 10.1111/j.1755-0998.2009.02778.x

TECHNICAL REVIEW

A practical guide to methods of parentage analysis

ADAM G. JONES, CLAYTON M. SMALL, KIMBERLY A. PACZOLT and NICHOLAS L. RATTERMAN
Department of Biology, 3258 TAMU, Texas A&M University, College Station, TX 77843, USA

Abstract

The use of molecular techniques for parentage analysis has been a booming science for over a decade. The most important technological breakthrough was the introduction of microsatellite markers to molecular ecology, an advance that was accompanied by a proliferation and refinement of statistical techniques for the analysis of parentage data. Over the last several years, we have seen steady progress in a number of areas related to parentage analysis, and the prospects for successful studies continue to improve. Here, we provide an updated guide for scientists interested in embarking on parentage analysis in natural or artificial populations of organisms, with a particular focus on computer software packages that implement various methods of analysis. Our survey of the literature shows that there are a few established methods that perform extremely well in the analysis of most types of parentage studies. However, particular experimental designs or study systems can benefit from some of the less well-known computer packages available. Overall, we find that parentage analysis is feasible and satisfying in most systems, and we try to provide a simple roadmap to help other scientists navigate the confusing topography of statistical techniques.

Keywords: exclusion, fractional allocation, microsatellites, parentage assignment, parental reconstruction, paternity

Received 16 May 2009; revision received 7 August 2009; accepted 1 September 2009

Introduction

Parentage analysis is a cornerstone of research in molecular ecology. Patterns of parentage play a central role in the study of diverse ecological and evolutionary topics, such as sexual selection (Yezerinac *et al.* 1995; Jones & Avise 1997a,b), patterns of dispersal and recruitment (Dow & Ashley 1996; Hardesty *et al.* 2006), estimation of quantitative genetic parameters (Kruuk *et al.* 2000; Garant & Kruuk 2005) and conservation biology (Haig 1998; Planes *et al.* 2009). However, parentage analysis is still a relatively young discipline. The molecular ecology revolution arguably started in the late 1960s with the advent of allozyme electrophoresis (Hubby & Lewontin 1966). In the early days of molecular ecology, parentage analysis was nearly impossible, with some notable exceptions (Ellstrand 1984; Gowaty & Karlin 1984), because of the low levels of polymorphism characteristic of protein-based markers. The study of parentage started to gain steam with the discovery that DNA probes could be used in humans and other organisms to reveal variation at minisatellite loci, a technique known as 'DNA fingerprinting' (Jeffreys *et al.* 1985). This multi-locus DNA fingerprinting approach was rapidly adopted by avian behavioural ecologists (Burke & Bruford 1987), resulting in tremendous insights into extra-pair mating (Westneat & Sherman 1997; Petrie & Kempenaers 1998). However, technical and statistical limitations slowed the spread of DNA fingerprinting applications outside of birds and mammals. Several years after the development of DNA fingerprinting, the discovery of microsatellite markers (Tautz 1989), also known as simple sequence repeats, blew the taxonomic doors wide open and started the current golden age of parentage analysis.

The introduction of microsatellite markers resulted in a complete parentage analysis overhaul, because they were the first easily assayable single-locus, codominant,

Correspondence: Adam G. Jones, Fax: (979) 845 2891;
E-mail: agjones@tamu.edu

© 2009 Blackwell Publishing Ltd

hypervariable markers (Avise 2004; Pemberton 2009). Even though multi-locus DNA fingerprinting was relatively simple methodologically, the banding patterns were extremely difficult to handle from a statistical standpoint, so parentage analysis had to be calibrated by the expected number of shared bands among different classes of relatives for any particular population or species. Microsatellites, on the other hand, followed the rules of Mendelian segregation, and key parts of the statistical theory related to parentage analysis had already been developed (Thompson 1975, 1976a; Meagher 1986; Meagher & Thompson 1986; Devlin *et al.* 1988). Thus, the field of parentage analysis was intellectually prepared for microsatellites, and plenty of empirical questions were waiting for answers. In this review, we explore how far we have come since the early days of microsatellite-based parentage analysis by answering a few key questions. For example, are microsatellites still the marker of choice? What kind of sampling strategy produces the best results? What type of parentage analysis technique is appropriate for a given system? And what are the remaining challenges in parentage analysis?

Types of parentage analysis

In a previous review, Jones & Ardren (2003) categorized parentage analysis techniques into four categories: exclusion, categorical allocation, fractional allocation, and parental reconstruction. In the last 6 years, a number of new techniques have been developed and each category of techniques could warrant an entire review article of its own. Thus, rather than belabouring the details of the methods, we describe the statistical approaches in a nontechnical way and refer the reader to specific articles that describe the nuts and bolts of each technique. In this article, we focus on the practical side of parentage analysis; so the main goal is to review techniques that have been implemented in computer software packages that are accessible to beginning researchers in the field of molecular ecology. Over the last several years, noteworthy improvements in methods have resulted in the addition of two new categories of statistical techniques that can now be fruitfully applied in the context of parentage studies. We call these two categories 'full probability parentage analysis' (Hadfield *et al.* 2006) and 'sibship reconstruction' (Thomas & Hill 2002). Here, we briefly describe the six categories of parentage analysis techniques and give shorter definitions in Box 1.

Exclusion

The logic behind exclusion is remarkably simple. Given the rules of Mendelian inheritance for diploid organisms, a parent and an offspring will share at least one allele per locus for a codominant marker (Chakraborty *et al.* 1974). If a putative parent fails to share an allele with the offspring of interest, then that parent can be eliminated from consideration as a true parent. However, despite the simplicity of the exclusion approach, some pitfalls are possible. First, any marker characteristic that prevents inheritance from appearing strictly Mendelian to the observer could result in false exclusions of true parents. Microsatellite markers are especially vulnerable to these types of phenomena. For example, a germ-line mutation will result in an offspring with an allele that is not present in the parent (Ellegren 2004; Eckert & Hile 2009), and null alleles (i.e., nonamplifying alleles) can result in a true parent and its offspring appearing homozygous for different alleles (Pemberton *et al.* 1995; Dakin & Avise 2004). Scoring errors result in the same type of problem (van Oosterhout *et al.* 2004; Hoffman & Amos 2005). Second, complete exclusion may be difficult if the population under consideration consists of a large number of offspring and putative parents. Even with highly polymorphic markers, it may be prohibitively expensive to genotype enough loci to achieve exclusion of all but the true parent for every offspring in the population. The solution to the first problem actually compounds the second and vice versa. To accommodate genotyping errors and mutations, most exclusion studies require mismatches at two loci before an exclusion is considered valid. This solution is a good one, as long as mutations and scoring errors are rare, but its successful implementation requires more powerful markers than would be required in the absence of scoring errors and mutation. Despite these considerations, complete exclusion is the gold standard of parentage studies; every parentage study is implicitly striving towards this goal.

Categorical allocation

Categorical allocation is the most commonly used method of parentage analysis. As in the case of exclusion, this approach requires at least one focal offspring and a set of candidate parents. Categorical allocation was developed as an answer to situations in which complete exclusion may not be feasible (Meagher & Thompson 1986). Since then, the categorical allocation approach has been updated and refined, resulting in a very useful technique for parentage assignment (Marshall *et al.* 1998; Gerber *et al.* 2000). The most obvious benefit of this approach is that it provides a method to choose the single most likely parent from a group of nonexcluded putative parents. The logic stems from the observation that different parental genotypes may differ in their probability of

© 2009 Blackwell Publishing Ltd

Box 1: Six approaches to parentage analysis

Exclusion – The exclusion method takes advantage of the fact that in diploid, sexually reproducing organisms, each parent shares at least one allele per locus with each of its offspring. In this approach, the genotypes of candidate parents are compared with that of a focal offspring. Any candidate parent who fails to share at least one allele with the offspring at any locus is eliminated from consideration. In practice, most exclusion studies actually require at least two mismatching loci between the candidate and the offspring to account for typing errors or mutations.

Categorical Allocation – If complete exclusion is impossible, then a parentage allocation approach (also known as parentage assignment) can be used to choose among the remaining nonexcluded candidate parents. In categorical assignment, the entire offspring is assigned to the candidate parent with the highest likelihood or posterior probability of being the true parent. Categorical assignment approaches can handle scoring errors or mutations and can include methods for determining confidence in parentage assignment.

Fractional Allocation – In the fractional allocation approach, likelihoods or posterior probabilities are determined in the same way as in the categorical assignment methods. Each offspring is then assigned partially to each of the nonexcluded candidate parents on the basis of their relative likelihoods of parentage. Even though a fractional assignment has no biological meaning, from a statistical standpoint, this approach may have better properties than categorical allocation.

Full Probability Parentage Analysis – The full probability approach estimates patterns of parentage in a modelling framework. Many different models are possible, but this approach has the potential to estimate simultaneously patterns of parentage and other population-level variables of interest. This approach makes better use of the data by incorporating any uncertainty in the parentage analysis into the estimation of the variables of interest.

Parental Reconstruction – The parental reconstruction technique uses the genotypes of offspring in full- or half-sib families to reconstruct parental genotypes. For full- or half-sib progeny arrays, all of the offspring will share at least one parent. The genotype of the shared parent may be available from the sampling scheme or can be reconstructed by identifying a pair of alleles, for which every offspring inherited at least one of the members of the pair. The genotypes of the unknown parents can be determined by examining associations of alleles originating from the unknown parents across loci. Available techniques are based on parsimony (i.e. assuming the minimum number of parents), maximum likelihood or Bayesian approaches. Once the genotypes are reconstructed, they can be compared with the genotypes of candidate parents to assign parentage.

Sibship Reconstruction – If no parents are available and known groups of full- or half-sibs cannot be sampled, then sibship reconstruction is the last resort in the realm of parentage analysis. This technique requires a sample of individuals, some of which are full- or half-sibs. The algorithms use patterns of relatedness or maximum likelihood techniques to group individuals into different classes of relationship, often full-siblings, half-siblings and unrelated individuals. Once half-sib or full-sib groups are identified by these approaches, the parental genotypes can be reconstructed and used for parentage analysis.

having produced the genotype of the focal offspring (Meagher & Thompson 1986). Most current categorical allocation approaches use a likelihood approach, but a Bayesian approach also can be used (Box 2). Either approach boils down to Mendelian transition probabilities (Marshall *et al.* 1998), which describe the probability of obtaining a certain offspring genotype given the proposed parental genotypes. For example, if an offspring has an unknown father and paternal allele '120', a male homozygous for the allele (i.e. with genotype 120/120) is more likely to have produced that allele than a male heterozygous for the allele (e.g. genotype 120/136). Clear descriptions of the full logic behind the likelihood approach are presented by Meagher & Thompson (1986) and Marshall *et al.* (1998). Nielsen *et al.* (2001) provide an excellent description of the calculation of Bayesian posterior probabilities in the context of parentage analysis. A positive feature of categorical allocation is that the use of likelihoods or posterior probabilities results in a framework in which scoring errors or mutations can be accommodated very easily by modifying the transition probabilities accordingly (Marshall *et al.* 1998; Kalinowski *et al.* 2007). Much like strict exclusion, the categorical allocation approach can be applied when one parent is known *a priori* for the focal offspring or when neither parent is known. In addition, this approach can be used to assign single parents or parent pairs. These different applications require different likelihood equations (Marshall *et al.* 1998), but they are all slight variations on the same theme. Overall, categorical allocation is a powerful, flexible approach to parentage analysis that has proven its worth in countless studies.

© 2009 Blackwell Publishing Ltd

Box 2. Assignment by likelihood or posterior probabilities

One possible division between approaches to parentage separates likelihood-based methods and Bayesian methods, although the distinction is relatively minor. Nevertheless, possible approaches to choose the best candidate parent from a group of nonexcluded individuals are worth discussing. Here, we assume that the mother is known and the goal is to choose the correct father from a pool of candidate males (which may or may not include the father). The equations are easily generalized to the cases where neither parent is known and either a single parent or parent pairs are the targets of assignment.

The likelihood approach uses a likelihood ratio to ask whether the current candidate male under consideration is more likely to be the parent of the offspring than a male chosen at random from the population. Here, we follow the logic of Marshall *et al.* (1998), which was based on the work of Meagher (1986). The likelihood ratio is the probability of the data given hypothesis one (H_1) divided by the probability of the data given hypothesis two (H_2), where H_1 is the case in which the candidate male is the true father and H_2 is the case in which the candidate male is unrelated to the offspring in question. The likelihood ratio is then given by

$$L(H_1, H_2|g_\mathrm{m}, g_\mathrm{f}, g_\mathrm{o}) = \frac{T(g_\mathrm{o}|g_\mathrm{m}, g_\mathrm{f})P(g_\mathrm{m})P(g_\mathrm{f})}{T(g_\mathrm{o}|g_\mathrm{m})P(g_\mathrm{m})P(g_\mathrm{f})} = \frac{T(g_\mathrm{o}|g_\mathrm{m}, g_\mathrm{f})}{T(g_\mathrm{o}|g_\mathrm{m})},$$

where the numerator represents the likelihood under H_1 and the denominator gives the likelihood under H_2 (Marshall *et al.* 1998). The genotypes of the mother, alleged father (i.e., candidate male) and offspring are given by g_m, g_f and g_o respectively. The transition probabilities, $T(g_\mathrm{o} \mid g_\mathrm{m}, g_\mathrm{f})$ and $T(g_o \mid g_m)$, represent the Mendelian probabilities of obtaining the offspring genotype given the mother and alleged father's genotypes or just the mother's genotype. These transition probabilities are easily derived, and an exhaustive list is given by Marshall *et al.* (1998). Finally, the expected frequencies of the maternal and alleged father genotypes are given by $P(g_\mathrm{m})$ and $P(g_\mathrm{f})$. The likelihood ratio is calculated on a per-locus basis and multiplied across independent loci. The natural logarithm of the multilocus likelihood ratio is called the LOD score (Meagher 1986). This simple likelihood ratio statistic serves as the basis for all likelihood-based categorical assignment techniques.

The alternative to a likelihood ratio is a posterior probability. The logic behind this approach is described well by Nielsen *et al.* (2001), so we present their equation here. We allow $\mathbf{O}_i$, $\mathbf{M}_i$ and $\mathbf{F}_i$ to represent the multilocus genotypes of offspring i, known mother i, and alleged father i. Furthermore, $\mathbf{A}$ is a matrix of allele frequencies for all loci, and we assume to have sampled n males from the total N males present in the breeding population. If $I_k(i)$ is the event that the potential father k is the actual father of offspring i, then

$$P(I_k(i)|\mathbf{M}_i, \mathbf{F}, \mathbf{A}, N) = \frac{P(\mathbf{O}_i|\mathbf{M}_i, \mathbf{F}_k)}{\sum_{j=1}^{n} P(\mathbf{O}_i|\mathbf{M}_i, \mathbf{F}_j) + (N-n)P(\mathbf{O}_i|\mathbf{M}_i, \mathbf{A})}.$$

The numerator of this equation is the probability of obtaining the offspring genotype given the known maternal genotype and the genotype of the candidate male (Nielsen *et al.* 2001), and this probability can be determined by assuming linkage equilibrium and Mendelian segregation (Thompson 1975). The term on the left in the denominator is the combined probability of paternity for all of the other sampled males, whereas the term on the right takes into account unsampled males. This equation has the advantage that it explicitly shows the dependence of assignment techniques on knowledge of the proportion of potential fathers sampled. In addition, the posterior probability approach includes information from all males in the population. The posterior probability approach, like the likelihood approach, can be used for categorical allocation or fractional allocation. However, posterior probabilities have the distinct advantage that they can be expanded very naturally to accommodate prior information or the estimation of other population-level variables. Consequently, this posterior probability approach serves as the basic framework for full probability parentage analysis (Box 3).

Fractional allocation

Given the apparent strengths of categorical allocation, why not stop here? The answer to this question is that different parentage studies have different goals and other techniques may outperform categorical allocation in certain contexts. Fractional allocation is similar to categorical allocation in many ways. The main difference is that categorical allocation assigns the entire offspring to the most likely male, whereas fractional allocation assigns a given offspring partially to each nonexcluded candidate parent based on their relative likelihoods or posterior probabilities (Devlin *et al.* 1988; Nielsen *et al.* 2001; Hadfield *et al.* 2006). Thus, in categorical allocation,

© 2009 Blackwell Publishing Ltd

the number of offspring for a candidate parent must be an integer, but in fractional allocation a candidate parent may be assigned a noninteger number of offspring. At face value, then, it appears that fractional allocation will seldom arrive at the absolute truth as it can be shown from first principles that each parent must have an integer number of offspring. Under many circumstances, however, categorical allocation will also fail to arrive at the complete truth, and fractional allocation may possess better statistical properties for many problems that involve the estimation of population-level variables, such as the relative fitnesses of genotypic classes or variances in reproductive success (see Devlin *et al.* 1988; Nielsen *et al.* 2001). For fractional allocation, likelihoods and posterior probabilities are calculated in the same way as for categorical allocation (Box 2). However, while categorical allocation approaches rely mainly on likelihood equations (Marshall *et al.* 1998; Gerber *et al.* 2000; Duchesne *et al.* 2005), Bayesian posterior probabilities are more often used for fractional assignment (Nielsen *et al.* 2001; Hadfield *et al.* 2006). In practice, likelihoods and posterior probabilities are very similar, so the preference of one over the other for a particular application represents more of a historical bias than a statistical reality. Regardless, despite showing some early promise, fractional allocation is seldom used in empirical studies.

Full probability parentage analysis

Even though we are treating parentage analysis as if it is a worthy goal in its own right, the truth of the matter is that parentage analysis often is interesting only because it allows the researcher to infer something about a population-level process (Jones & Ardren 2003; Hadfield *et al.* 2006; Pemberton 2009). The full probability parentage analysis techniques embrace this point of view (Nielsen *et al.* 2001; Hadfield *et al.* 2006). This method estimates the population-level parameters of interest simultaneously with the parent-offspring relationships in a single modelling framework that interfaces very naturally with the fractional allocation techniques (Roeder *et al.* 1989; Smouse *et al.* 1999; Morgan & Conner 2001; Nielsen *et al.* 2001). In addition, a Bayesian approach also permits the inclusion of prior data, such as dominance rank or spatial location of candidate adults, increasing the prospects for successful assignment (Neff *et al.* 2001; Hadfield *et al.* 2006).

An important advantage of the full probability approach is that uncertainty in the parentage analysis is included in the estimation of the population-level variables of interest. In categorical assignment approaches, for example, the estimation of such variables is a two step process, in which the patterns of parentage are estimated first and then, given those patterns of parentage, the population-level variable of interest is estimated. Thus, during the estimation of the variable of interest, any uncertainty in parentage is ignored, possibly resulting in an elevated level of confidence in the estimate (Hadfield *et al.* 2006). The full probability approach includes any uncertainty in parentage assignments in the analysis, resulting in a more accurate assessment of confidence in estimates of variables of interest. In addition, the ability to include prior information opens the possibility to make better use of the available data (Neff *et al.* 2001). For example, categorical and fractional assignment techniques make the implicit assumption that (in the absence of genotypic data) all parents included in the analysis are equally likely to be the true parents of a given offspring. Full probability models can relax this assumption by taking into account relevant ecological information, such as territoriality, spatial location, breeding status and so forth.

Full probability parentage analysis approaches can take several forms, all of which are fairly involved. The basic approach is to specify a model of the probability of parentage that includes other explanatory variables in addition to the genotypes of parents and offspring (Box 3). The relationships between some of these variables and the probability of parentage could be known with certainty and this knowledge would affect prior probabilities of parentage for certain individuals in the population. Other variables could have unknown relationships with parentage and the estimation of these relationships would be a part of the analysis of the model (Box 3). For example, if something is known about dominance status, then dominant males might be expected to have a different share of paternity compared with subordinate males. However, the difference in reproductive success among the two groups may be unknown (Nielsen *et al.* 2001). The full probability approach allows the model to include the possibility that dominant males systematically produce a different number of offspring than subordinate males, while leaving the exact difference between the two groups a variable that can be estimated during the analysis of the genetic data (Nielsen *et al.* 2001; Hadfield *et al.* 2006). In principle, this approach can be expanded to accommodate almost any parameter of interest.

Full probability parentage analysis appears to be a promising approach, but it has yet to be widely embraced by molecular ecologists. Part of the problem is that the specification of the model is difficult, and it may require undocumented assumptions about the relationships between particular variables and mating success. For example, Hadfield *et al.* (2006) assumed that the relationship between a male's probability of paternity and his distance from an offspring can be modelled using an exponential function. If this assumption is incorrect, the parentage assignment could be compromised. For many taxa, there may be insufficient data to construct a well-supported full probability model. The other potential pit-

© 2009 Blackwell Publishing Ltd

Box 3. Full probability parentage analysis

A potentially powerful approach to parentage analysis is to estimate population-level variables of interest at the same time as the patterns of parentage. Here we give an intuitive description of the logic behind this type of approach without delving too deeply into the mathematics and jargon. This approach is a natural extension of the posterior probability approach to parentage allocation described in Box 2. The example that we will use to illustrate full probability models comes from Hadfield *et al.* (2006), who have developed a flexible, general framework for the implementation of this method.

The logic is most easily seen by considering the basis of the model given by Hadfield *et al.* (2006). This model of parentage in the Seychelles warbler includes dominance status of the females and distance between males and offspring. For each offspring, imagine a table with a row for each breeding female and a column for each breeding male. Each entry in the table is the probability that the corresponding male and female were the parents of the offspring in question. In the Seychelles warbler, the entries in this table can be given by the following equation (Hadfield *et al.* 2006, equation 2):

$$P(\mathbf{a}_m^{(i)} = j, \mathbf{a}_f^{(i)} = k) \propto \frac{\theta^{s_j}(1-\theta)^{1-s_j}}{\sum_{m=1}^{n_m} \theta^{s_m}(1-\theta)^{1-s_m}} \frac{e^{-\mathbf{d}_k^{(i)}\lambda}}{\sum_{f=1}^{n_f} e^{-\mathbf{d}_f^{(i)}\lambda}}$$
$$\prod_{l=1}^{L} \frac{P(\mathbf{O}_{l,i}|\mathbf{F}_{l,k}, \mathbf{M}_{l,j})}{\sum_{m=1}^{n_m}\sum_{f=1}^{n_f} P(\mathbf{O}_{l,i}|\mathbf{F}_{l,f}, \mathbf{M}_{l,m})}.$$

This expression gives the relative probability of female j and male k being the parents of offspring i. The first term is the dominance model for females, with s_j indicating whether the female is dominant or not (1 = dominant, 0 = subordinate) and θ as an unknown parameter that represents the ability of dominant mothers to outcompete subordinate mothers. The parameters n_m and n_f are simply the numbers of mothers and fathers in the population. The second term in the equation is the distance model, where $\mathbf{d}_k^{(i)}$ gives the distance of male k from offspring i, and λ is an unknown parameter that determines the rate at which probability of paternity drops off as a function of distance. The final term is the probability of parentage based on the genetic data, similar to the posterior probability in Box 2, with L representing the number of loci.

Given the model, the challenge is then to choose values of θ, λ, and a pair of parents for each offspring that maximizes the joint probability across all offspring given the data. As the solution involves explicit probabilities for all possible sets of parents for every offspring in the data set and the estimation of the parameters of interest requires integration across uncertainty in parentage assignments, this problem can be computationally intensive. One solution is to use a Markov Chain Monte Carlo approach (Hadfield *et al.* 2006), which produces point estimates for the parameters of interest as well as an indication of the level of statistical uncertainty for each estimate. Other approaches also may be feasible depending on the precise model.

fall with full probability models is that the practice of plugging data into a complex analytical technique that spits out an answer with confidence limits is potentially dangerous. Especially for parentage analysis, we advocate an approach in which researchers carefully examine the assignments and consider whether there are logical inconsistencies in the data set or not. Blind application of the full probability model could hide defects in the data set from the user. Thus, even if full probability models are the method of choice, we encourage this approach to be augmented by examination of the actual probabilities or likelihoods of parentage, which should be inspected by the researcher for inconsistencies or unexpected patterns. Any strange results should be subjected to additional scrutiny, possibly entailing the collection of additional genetic data.

One extremely important point to keep in mind in the interpretation of full probability models, or any other technique involving assignment, is that all of the techniques we have described so far will converge on the same answer if the genetic data are sufficiently strong. If parents can be identified for all offspring in the data set with certainty, then any parameter of interest can be justifiably estimated from the parentage assignments. In a perfect world, exclusion, categorical assignment, fractional assignment and full probability models would all produce the same answer. From a practical standpoint, a good approach might be to analyse the data using at least

© 2009 Blackwell Publishing Ltd

two different methods and compare the results across the different analyses. Any large discrepancies may indicate the need for additional genetic data.

Parental reconstruction

Parental reconstruction techniques take advantage of the fact that we sometimes have prior knowledge that certain groups of individuals probably originated from the same family. For example, we could collect an amphibian egg mass for which all of the offspring had the same mother (Myers & Zamudio 2004), a male defending a nest full of eggs (Jones *et al.* 1998b; Neff 2001), or a group of offspring developing inside the body of a parent (Tatarenkov *et al.* 2008; Mobley & Jones 2009). In these and similar cases, when a progeny array is known to include only half- and full-sibs, the genotypes of the parents can be reconstructed from the genotypes of the progeny (Jones & Avise 1997b; DeWoody *et al.* 2000a,b). If one parent is known, then the alleles of the unknown parents can be determined by subtracting the known parent's alleles from the offspring genotypes. The number of alleles from unknown parents per locus in the progeny array provides some indication of the number of unknown parents. In addition, associations of alleles across loci can allow the reconstruction of the unknown parents' genotypes (Jones 2001, 2005). For example, if the alleles 122 and 134 from an unknown parent are segregating at a locus in the progeny array and they are always associated with alleles 222 and 230 from an unknown parent at a second locus, then logic would suggest that the genotype of one of the unknown parents consisted of 122/134 at the first locus and 222/230 at the second. For progeny arrays with no known parents, the problem is not much more difficult. Provided the progeny array includes only full- and half-sibs, all offspring will share at least one parent. For any given locus, the genotype of the shared parent can be inferred by identifying a pair of alleles for which every offspring has at least one allele from the pair. Once the genotype of the shared parent is identified, the inference of the other parents follows the logic above.

Parental reconstruction can work extremely well if markers are sufficiently polymorphic. Several approaches have been implemented. One approach is a brute-force exhaustive algorithm that tests every possible genotype consistent with the alleles in the progeny array to identify the minimum number of parents necessary to explain the array and their genotypes (Jones 2001, 2005). Other approaches use Bayesian or maximum-likelihood techniques to identify likely partitions in the progeny arrays between half-sib groups originating from different unknown parents (Emery *et al.* 2001; Wang 2004). As in the case of assignment techniques, these different parental reconstruction techniques all converge on the same answer when the molecular markers are sufficiently powerful.

Parental reconstruction has several favourable characteristics that allow it to complement other parentage analysis techniques nicely. For example, the genotypes recovered often are sufficiently rare that parents can be matched to progeny arrays with extremely high confidence (Jones *et al.* 2002). In addition, parental genotypes can be reconstructed from progeny arrays in the absence of a pool of candidate parents (Emery *et al.* 2001; Jones 2005), allowing the mating behaviour of an uncollected gender to be inferred, provided the sample of progeny is sufficiently complete. Finally, the genotypes in the family groups provide an internal mechanism for identifying suspicious genotypes that may be the result of scoring errors, mutations or null alleles (Wang 2004). As the progeny are known to be relatives, any unexpected genotypes can be viewed askance and subjected to greater scrutiny. Thus, even though not all of these methods accommodate genotyping error, the internal checks within the progeny array make it possible to eliminate the vast majority of the errors from the data set. These favourable characteristics lead us to suggest that whenever possible, scientists should endeavour to sample in a way that retains any information about family structure in the offspring of interest.

Parental reconstruction, while extremely useful for some purposes, does have some drawbacks. First, it requires highly polymorphic markers, and many systems may not have sufficiently polymorphic loci for the technique to work well. Second, parental reconstruction is much more effective on large progeny arrays. If an unknown parent has fewer than 8–10 offspring in the progeny array, the prospects for successful reconstruction diminish considerably. This latter constraint also prevents the method from being effective in families with an extremely large number of parents per progeny array (i.e. more than about half a dozen). The rationale behind this limitation can be seen by considering binomial probabilities. When the number of offspring from a particular parent in a progeny array is less than about six, there is a reasonable probability that only one of a heterozygous parent's two alleles at a locus will segregate in the progeny array, a situation that would preclude correct reconstruction of the parental genotype. Thus, parental reconstruction has enjoyed good success in organisms with large brood sizes, but has been less useful in species with small families, such as birds or mammals.

Sibship reconstruction

The final category of techniques that we discuss, sibship reconstruction, used to be on the fringe of parentage anal-

© 2009 Blackwell Publishing Ltd

ysis (Blouin 2003; Jones & Ardren 2003), but the algorithms have been improving and now can provide reconstructed parental genotypes or use candidate genotypes to guide the sibship reconstruction procedure (Almudevar & Field 1999; Thomas & Hill 2002; Ashley *et al.* 2008). Sibship reconstruction comes into play when a group of offspring can be collected from the population, but family groups cannot be identified *a priori* even though the sample is known to contain some full- and half-sibs. The underlying idea is to use genotypic data of the individuals in the sample to partition individuals into groups of full siblings or groups of full siblings and half siblings.

Methods for sibship reconstruction have recently been reviewed elsewhere (Ashley *et al.* 2008), so we give only a brief overview here. Current techniques for sibship reconstruction fall into two major categories. The first category includes likelihood-based methods, in which the algorithm attempts to partition the sampled individuals into sibling groups in a way that maximizes the probability of the data (Smith *et al.* 2001; Thomas & Hill 2002; Wang 2004). For example, one way of computing likelihoods is to determine the probabilities of all half-sib families, some of which are nested within full-sib families, given the rules of Mendelian inheritance across all possible parental genotypes that could exist in the population assuming Hardy-Weinberg equilibrium (Wang 2004). A second category of sibship reconstruction techniques includes combinatorial approaches (Almudevar & Field 1999; Berger-Wolf *et al.* 2007; Ashley *et al.* 2008), which take advantage of a strong focus on Mendelian segregation to retain sibling groups that adhere to Mendel's laws. The distinction between combinatorial and maximum likelihood approaches is blurred, however, as both take advantage of Mendelian segregation, and essentially all implementations of these methods are sufficiently computationally challenging that stochastic optimization techniques are required to obtain a solution in a timeframe relevant to a typical human lifespan.

Sibship reconstruction can be useful in the context of parentage analysis when a large group of offspring can be collected, but they are not associated with any particular parent and not in family groups. If a pool of candidate parents is available, then an assignment technique can be used, with sibship reconstruction serving as a complementary approach. If candidate parents are not available, then sibship reconstruction could allow some inference of patterns of parentage through the comparison of reconstructed genotypes.

A parentage analysis roadmap

Assuming that parentage analysis is the right approach for the empirical question at hand, one of the most important questions concerns experimental design. Now that we have seen the various methods available, what kinds of considerations are necessary to increase the probability of success in parentage analysis? How can we collect samples to ensure that the best statistical method is available for our data set? Clearly, the sampling design is one of the most important factors determining the efficacy of a study (Jones & Ardren 2003; Pemberton 2009), but the unfortunate reality is that many systems are limited because of various unfavourable biological properties of the organisms. Another major consideration is the choice of molecular markers. For most parentage issues, we would like to have many loci with extremely high levels of polymorphism per locus, but again the biological reality might be that a particular study system has less than adequate markers. Regardless, successful parentage analysis requires an investment in reliable markers and suitable quality control (Jones & Ardren 2003; Pemberton 2009).

The importance of good sampling

In Fig. 1, we present a flow chart that indicates the types of analyses that will be available given various characteristics of a system. We start with the question of whether a pool of candidate parents can be identified and sampled or not. For example, if individuals of one gender can be collected with their offspring, is there an accessible group of candidate parents that likely mated with the known parents? If candidates can be identified, then all methods of parentage analysis are a possibility. In the absence of candidate parents, the options are fairly limited. If half- or full-sib groups can be identified and these groups are relatively large, then parental reconstruction remains an option. If only small family groups can be collected, then it still may be possible to look for evidence of multiple mating within family groups, an approach that we do not review here, but that can be tackled with parental reconstruction or sibship reconstruction methods among others (DeWoody *et al.* 2000a,b; ; Neff *et al.* 2002; Sefc & Koblmüller 2009). If offspring cannot be collected in family groups, however, sibship reconstruction may be the only suitable method. If sibship reconstruction or parental reconstruction succeeds, then the reconstructed parental genotypes can be compared to learn something about mating patterns of breeders, including adults for whom tissue samples could not be obtained.

Parentage analysis is normally applied to systems in which candidate parents can be collected, so most techniques assume that there will be a sample of adult genotypes. From a parentage analysis standpoint, the ideal study would involve large family groups of progeny col-

© 2009 Blackwell Publishing Ltd

lected with a shared parent, as would be possible for fruits collected from a plant (Teixeira & Bernasconi 2007) or pregnant female live-bearing fishes (Neff *et al.* 2008), plus a complete sample of all potential breeders in the population. Such a complete sample can be extremely difficult to obtain in natural populations, but it allows all parentage analysis techniques to be applied, resulting in excellent prospects for success. For this type of sample, our approach would be to use parental reconstruction to determine genotypes of the unknown parents, match them to the pool of candidate parents and verify the matches using an assignment or exclusion approach (Jones *et al.* 2002). In many systems, however, large groups of related progeny do not occur, so parental reconstruction would not apply. In addition, depending on the goals of the study, it may require a prohibitively large amount of genotyping to assay more than a handful of offspring per family. Nevertheless, the prospects for parentage inference are still good as long as candidate parents are available, because exclusion and assignment would still be options. In all of these cases, particularly if the parentage analysis techniques indicate <95–100% confidence in assignments, the full probability models should be considered as an option for estimating population-level variables of interest.

The bottom line is that the most difficult part of a parentage study is the sampling of specimens. Unfortunately, it also is the most important part of a study. Every effort should be made to ensure that samples from the field are as complete as possible. One point that we should emphasize is that all the above techniques are more powerful when one parent is known with certainty for each offspring than when neither parent is known, so the ability to collect offspring with a parent facilitates parentage analysis. If an adequate sampling scheme is not possible for an organism of interest, scientists should seriously consider whether or not their question of interest can be addressed in a more tractable system.

Molecular markers

In principle, any variable genetic marker that is stably transmitted from parents to offspring can be used for parentage analysis. However, certain types of markers are better suited to this application than others (DeWoody 2005; Pemberton 2009). Successful parentage analysis requires either highly polymorphic markers or a very large number of markers with low to moderate levels of polymorphism. Even though many kinds of markers have

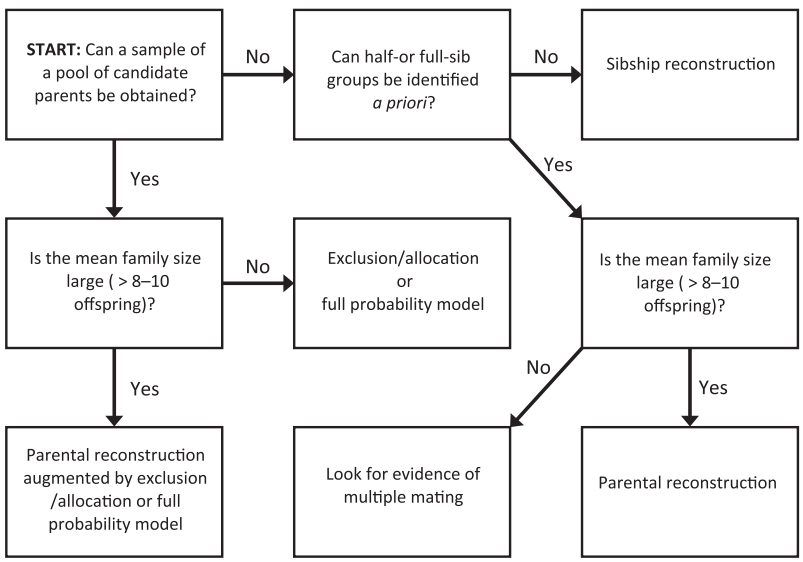

Fig. 1 A flow chart showing the different parentage analysis techniques available for various types of sampling schemes. Start in the upper left corner and answer the questions to see which analysis techniques are available in your study system. The best case scenario is shown in the lower left corner: a system in which large families can be collected with a pool of candidate parents. This type of sample allows all parentage analysis techniques to be applied. As samples depart from this ideal, the options become fewer and fewer, until parentage analysis is no longer even possible (lower middle). In this worst case scenario, it may still be possible to look for evidence of multiple paternity or multiple maternity depending on the nature of the sample. Techniques for detecting multiple mating are many and varied, and they are beyond the scope of this review. However, parental reconstruction techniques, even when applied to a few offspring, typically will be able to distinguish progeny arrays arising from single mating from those arising from multiple mating. Even though it is not addressed in this figure, parentage analysis is more efficient when one parent is known with certainty (e.g. when mothers or fathers can be collected with offspring that are known to be theirs).

© 2009 Blackwell Publishing Ltd

been used in parentage analysis over the years, a few techniques have distinguished themselves as the most useful.

The overall best marker for parentage analysis in almost every situation is the microsatellite (Pemberton 2009). These markers are highly polymorphic, co-dominant, PCR-based and repeatable. Even though the development of microsatellites is somewhat involved, this approach is feasible for most types of organisms. Our advice for anyone embarking on parentage analysis is to invest in the development of a suite of reliable microsatellites. Most microsatellites used for parentage analysis are di-, tri-, or tetra-nucleotide repeats, meaning that the repeated motif consists of either two, three or four base pairs. Tetra-nucleotide microsatellites in particular are appealing, because different alleles are usually separated by four base pairs, making them more easily separable on a gel, and tetra-nucleotides compared with di-nucleotides suffer from less 'stuttering' (Walsh *et al.* 1996), which results in bands slightly larger or smaller than the true allele on the gel.

While microsatellites will be the marker of choice for most studies, two other types of markers, single nucleotide polymorphisms (SNPs) and amplified fragment length polymorphisms (AFLPs), are viable choices. We will leave the details of the methods to other reviews (Morin *et al.* 2004; Meudt & Clarke 2007), but both of these techniques take a different approach to the problem than microsatellites. While microsatellites usually operate in the realm of few highly polymorphic loci, AFLPs and SNPs get the job carried out with very many loci, each of which has low levels of polymorphism. Usually, each technique reveals only two alleles per locus. In the case of SNPs, the alleles are codominant, whereas AFLPs are dominant markers. In other words, AFLPs do not allow the heterozygote to be distinguished from one of the homozygous classes. The AFLP technique has been around for over a decade, but for some reason has been used mainly for parentage analysis in plants (Gerber *et al.* 2000; Sezen *et al.* 2009). These markers have the advantage that they require very little development, as the technique can be applied to almost any organism by using a commercially available kit. In contrast, SNP markers have been gaining popularity with the proliferation of genomics data. They depend on known DNA sequences in which particular nucleotide positions have been shown to be polymorphic. Thus, SNPs are more easily obtained for model organisms, and for those organisms many hundreds of loci may be available. On a per-locus basis, SNPs are more easily assayed than microsatellites. Some researchers have predicted that SNPs will be the marker of choice for parentage analysis in the future (Anderson & Garza 2006). However, one consideration for AFLPs or SNPs is that some approaches to data analysis are precluded by the low per-locus levels of polymorphism. For example, existing techniques of parental reconstruction for a particular brood would perform poorly with these markers, because for a di-allelic locus, a single parent heterozygous at all loci is compatible with all offspring. In other words, for a di-allelic locus, when one parent is known with certainty, the minimum number of unknown parents for a progeny array will always be one. A similar type of argument suggests that sibship reconstruction will not work well with AFLP or SNP data.

These considerations lead us to agree with Glaubitz *et al.* (2003) and predict that microsatellites will remain the marker of choice for parentage analysis for quite some time. We expect SNPs to increase in popularity for exclusion or parentage assignment in model systems or heavily managed systems for which genomic data are available, but SNPs probably will not be widely used by garden variety molecular ecologists. We also predict that AFLPs will diminish in popularity, but they will still be useful for quick and dirty parentage studies for species in which development of microsatellites is not warranted. We would like to add one caveat, which is that it may be possible to increase the utility of SNPs by using tightly linked loci in an experimental design that allows linkage phase to be assessed (Jones *et al.* 2009). In this situation, the linked SNPs become a sort of 'super locus', potentially with many alleles, provided the rate of recombination is low enough that haplotypes are stably inherited. This philosophy also could be applied to microsatellite loci in species with low levels of polymorphism to increase the utility of the markers (Jones *et al.* 1998a,b; Estoup *et al.* 1999). This 'super-locus' situation is the only one in which linked loci are recommended for parentage analysis. Except for this special case, loci that are in linkage disequilibrium should be avoided because the statistics of parentage analysis typically assume independence among loci (Jones & Ardren 2003; also see Devlin *et al.* 1988).

A brief comment on exclusion probabilities

Studies of parentage often report exclusion probabilities for their molecular markers. These values represent the probability that an unrelated candidate parent (i.e. a genotype chosen at random from the population) will be eliminated from consideration as a true parent by the locus in question (Chakraborty *et al.* 1988). Three different exclusion probabilities are typically calculated: one parent known, neither parent known and parent pairs known (Jamieson & Taylor 1997). Exclusion probabilities are not especially useful because they assume an absence of mutations and scoring errors. They also are derived for single-offspring parentage tests, so they do not correct for experiment-wide error that arises from the hundreds of comparisons comprising a typical parentage study. Some progress has been made on deriving more informa-

© 2009 Blackwell Publishing Ltd

tive exclusion probabilities (Wang 2007; Baruch & Weller 2008) and developing other criteria for choosing loci (Matson *et al.* 2008). However, the traditional exclusion probabilities are easy to calculate and they provide a comparative measure of marker information content, so we recommend reporting exclusion probabilities for each marker along with other summary statistics such as allelic richness or heterozygosity. The exclusion probability should not be used as a measure of confidence in a parentage analysis, as this goal can be better accomplished by using methods that have been implemented in various software packages. In addition, many of these programs provide simulation-based approaches for assessing the expected power of a set of markers to resolve parentage in particular systems (Marshall *et al.* 1998; Jones 2001; Duchesne *et al.* 2005).

Other considerations

Genotyping errors, mutations and null alleles

With a good sampling design and a nice battery of molecular markers in hand, the stage is set for a successful study, but a few other details require attention. Characteristics of the markers and the fact that they are analysed by fallible humans can result in inconsistencies that present problems for parentage analysis. The most important class of inconsistencies concerns genotyping errors and mutations. Genotyping errors occur when a genotype is misread, fails to amplify, or spuriously produces a misleading result. Mutations on the other hand are a real biological phenomenon in which the allele inherited by the offspring changed in some way from the allele present in the parent. In the very large studies typical of parentage analysis, especially for highly variable markers, both types of problems arise and result in apparent incompatibilities between true parents and their offspring. Care must be taken to accommodate such errors in the parentage analysis (Marshall *et al.* 1998; Hoffman & Amos 2005). In exclusion or parental reconstruction, for example, a common practice is to exclude parents or invoke an additional parent only if the result can be verified by at least one additional locus. However, such an approach may be overly conservative, possibly resulting in many incorrect inclusions, so care must be taken to ensure that sufficient power is available to tolerate such a conservative approach. On the other hand, categorical or fractional assignment, full probability models and sibship reconstruction approaches can accommodate error by building a model of error into the calculation of likelihoods or posterior probabilities (Wang 2004; Koch *et al.* 2008). Very little has changed with respect to how genotyping errors and mutations are handled in most parentage analyses in the last 6 years, so we refer the reader to Jones & Ardren (2003) for more information.

Another potentially major problem in parentage analyses stems from nonamplifying alleles, otherwise known as null alleles (Dakin & Avise 2004). Such alleles can result in a mismatch between a parent and an offspring, but the mismatch will invariably involve apparently homozygous genotypes that are actually heterozygous for the null allele. For the most part, null alleles are not handled well by parentage analysis programs. However, they usually can be detected either as a departure from Hardy-Weinberg equilibrium at the null-bearing locus or as a non-Mendelian pattern of segregation in known family groups (Chakraborty *et al.* 1992; Brookfield 1996; Kalinowski & Taper 2006). As loci with null alleles are usually identifiable, they tend not to be a major problem for parentage analysis. Jones & Ardren (2003) provide a lengthier discussion of null alleles that is still valid.

The most noteworthy development related to genotyping error, mutations and null alleles over the last several years came from Wang (2004), who introduced a model for handling errors that includes two types of errors, allelic dropouts and stochastic errors. This model, which is similar to that of Kalinowski *et al.* (2006a,b), can accommodate null alleles, which are basically systematic allelic dropouts, as well as microsatellite mutations and scoring errors, which fall into the stochastic error category. The method also permits the rate of error to differ among loci. Wang's (2004) approach has been implemented in sibship reconstruction algorithms (Wang 2004) as well as in full probability parentage analysis (Koch *et al.* 2008). However, further testing will be necessary before we know how much of an increase in the accuracy of parentage analysis to expect from this way of handling errors and null alleles.

Family structure in the candidate parents

Sometimes the pool of candidate parents will include relatives of one another or of the offspring of interest. As most inference techniques assume that the candidate parents are unrelated to one another and to the offspring, except for parent-offspring relationships, the presence of family structure in the population can be problematic (Double *et al.* 1997; Olsen *et al.* 2001). This problem has been investigated in some detail in various studies (Marshall *et al.* 1998; Nielsen *et al.* 2001; Duchesne *et al.* 2008), and parentage analysis usually is not seriously impacted unless very close relatives of the offspring are included in the pool of candidate parents. Nonexcluded full siblings of the offspring, for example, actually can have higher likelihoods of parentage than the true parents (Thompson 1975, 1976a,b; Thompson & Meagher

© 2009 Blackwell Publishing Ltd

1987). If many close relatives of the offspring are likely to be in the pool of candidate parents, complete exclusion or parental reconstruction techniques, which are relatively insensitive to family structure in the population, may be necessary to reliably diagnose the true patterns of parentage (Jones & Ardren 2003).

Assessing confidence in parentage analysis

Perhaps the most important development in parentage analysis since the discovery of microsatellites has been the implementation of techniques capable of assessing confidence in the assignments. The first comprehensive approach, which remains the most popular, was developed by Marshall *et al.* (1998). Their error-handling approach has been updated recently (Kalinowski *et al.* 2007) for the computer program CERVUS 3.0, but the method remains otherwise unchanged. The approach is to simulate populations of breeders and their offspring given a user-specified rate of genotyping error and proportion of candidate parents sampled. Parentage is assigned on the basis of LOD scores (Box 2). A test statistic, Δ, is calculated for each assignment. For a given offspring, if only one candidate parent has a positive LOD score, Δ is the LOD score, but if two or more candidates have positive LOD scores, then Δ is the difference between the two highest LOD scores. The approach is to compare the distribution of Δ values for correct assignments in the simulation to that of false assignments, and to pick a critical value for Δ that gives a desired level of confidence in assignment. This critical value chosen from the simulation is then used to determine confidence in assignment for the actual analysis of the empirical data set. This approach not only yields an experiment-wide error rate for the assigned individuals, but it also indicates for each focal offspring, whether the parent is confidently assigned or not.

Even though the Marshall *et al.* (1998) approach is effective and popular, several alternatives also exist. For example, the use of Bayesian posterior probabilities is on the rise (Box 2). These methods give the posterior probability of each assignment on an offspring by offspring basis (Nielsen *et al.* 2001; Koch *et al.* 2008). Posterior probabilities seem to offer great promise, but some work remains to be carried out. First, algorithms should provide a posterior probability cut-off value that gives a desired level of experiment-wide error, as the Marshall *et al.* (1998) approach does. Second, the posterior probability approach should be compared with the likelihood approach by taking advantage of simulated data sets or experimental populations. Some work along these lines has been carried out (Koch *et al.* 2008), but the issue is far from complete resolution.

The third major approach to assess confidence for a parentage analysis algorithm is to use simulations to estimate an experiment-wide expected error rate without attributing confidence to any particular assignment (Jones 2001, 2005; Duchesne *et al.* 2005). This approach involves the simulation of data sets under user-specified parameters. These simulated data sets are then subjected to the parentage algorithm, and the number of success and failures gives a measure of the expected rate of success. Methods that use this approach should be employed only when experiment-wide confidence can be shown to be high.

We address the approaches used for confidence assessment in more detail for particular programs in the Appendix. Additional details regarding some of the older programs can be found in Jones & Ardren (2003). Programs that do not use one of the aforementioned approaches (i.e. simulation-based determination of critical values for a test statistic, posterior probabilities, or experiment-wide error rates determined by simulation) should be used with caution. For example, some programs use *ad hoc* methods or make no attempt to correct for experiment-wide error, making the interpretation of results difficult.

Two key variables requiring attention

The discussion of confidence in parentage assignment leads to the consideration of two variables in parentage analysis that are so important, especially for assignment techniques, that they warrant their own section. As the determination of confidence in parentage assignment is based on a model, an incorrectly parameterized model will result in incorrect estimation of confidence. From the work that has been conducted so far, the two most important user-supplied parameters seem to be the rate of genotyping error and the proportion of candidate parents sampled, both of which can be difficult to determine.

A complete, convincing study of parentage now usually will require some estimate of genotyping error (Hoffman & Amos 2005). Various techniques can be used to detect genotyping errors. The simplest situation occurs in parental reconstruction, where the known relatives serve as a reference (Wang 2004) and any unusual alleles can be investigated in detail or subjected to repeated genotyping. In the absence of progeny arrays, if one parent is known with certainty, then mismatches between the known parents and their offspring can provide a measure of the rate of genotyping error. Otherwise, genotyping error should be estimated by repeated typing of a subset of individuals (Hoffman & Amos 2005; Johnson & Haydon 2006). Merely knowing the error rate is not sufficient, however. Scientists should take steps to reduce errors as much as possible, because the presence of errors

© 2009 Blackwell Publishing Ltd

in the data set, even if the error rate is known, can substantially reduce the power of parentage analysis (Marshall *et al.* 1998).

The proportion of candidate parents sampled is the other key variable that can dramatically affect the probability of having a pleasing parentage analysis experience. Exclusion or parental reconstruction, given enough power and a data set with few errors, is relatively insensitive to the proportion of candidate parents sampled. For these techniques, offspring with unsampled parents will simply have no compatible parents in the pool of candidates. With less powerful markers, however, assignment techniques come into play and the evaluation of confidence requires some knowledge of the number of unsampled candidate parents. Incorrect specification of this parameter will result in incorrect estimates of confidence in assignment, possibly to an extreme degree (Marshall *et al.* 1998; Nielsen *et al.* 2001; Oddou-Muratorio *et al.* 2003). Solutions to this problem include ecological estimates of the breeding population size via a mark-recapture study or estimates of the number of breeding individuals from the genetic data (Ramakrishnan *et al.* 2004). Another option that makes use of the genetic data is to use a full probability parentage approach in which the number of candidate parents is estimated along with the patterns of parentage (Nielsen *et al.* 2001; Signorovitch & Nielsen 2002; Koch *et al.* 2008). How well this approach actually performs is a subject for future testing. Regardless, scientists have an obligation to include uncertainty in breeding population size in their results by performing analyses under parameter values that span the range of possibilities for their study system (Koch *et al.* 2008).

One caveat regarding the importance of these two variables is that poor estimates usually will not affect the rank order of the candidate parents as far as relative likelihoods are concerned. Rather, the rank order of compatible parents will stay the same, but the confidence in assignment will be incorrectly calculated. Thus, if the rate of genotyping error or the number of candidate parents cannot be determined, additional genotyping could compensate for any error introduced by the lack of knowledge of these variables. In other words, with more informative genotypic data, parentage analysis becomes less sensitive to error in estimating genotyping error and the number of candidate parents.

Launching into the software

We summarize the current software available for parentage analysis in Tables 1–3. Table 1 includes the software packages that we suggest as a starting point for anyone embarking on parentage analysis. The programs in Table 1 are for the most part user friendly and most have been used widely enough to demonstrate that they work well. A typical molecular ecologist could get by with just programs from Table 1. We have categorized the programs by what we perceived to be their main use, but we also indicate other possible applications. For example, CERVUS is designed for categorical assignment, but it also can be used for strict exclusion. Similarly, COLONY is a sibship reconstruction program designed to analyse many families simultaneously, but it can also assign parents if parental genotypes are specified and it performs parental reconstruction when applied to a single family (or multiple families at once). Nevertheless, our categorizations can be interpreted as recommendations. In other words, while COLONY can assign parents to offspring, we would not use it for this purpose over CERVUS.

In Table 2, we provide a list of 'niche programs' possessing unusual features that make them suitable for analysis of data sets that differ in particular ways from the norm. For example, FAMOZ is a categorical allocation program that can accommodate dominant markers, so it might be the program of choice for an AFLP data set. Other researchers have access to gender-linked markers or are working in haplo-diploid systems. These cases can be handled by NEWPAT for exclusion involving sex-linked markers, and MATESOFT, which estimates parental genotypes in haplo-diploids. A few other niche programs are listed in Table 2 as well. Finally, in Table 3, we provide a list of other programs that may be useful for the adventurous analyzer of parentage. Most of the programs in Table 3 perform functions that can be more effectively accomplished using a program from Table 1. No doubt some of the authors of these programs will feel that their program should be moved up to Table 1 or Table 2, so we now pass the ball into their court and ask for empirical data showing that their program outperforms other programs in some important way.

In the Appendix, we provide a brief description of each program listed in Tables 1–3. These descriptions are short, so they do not capture the full suite of capabilities for some of the programs. However, we hope that the Appendix will provide a starting place for individuals interested in finding more information about any of these parentage programs.

The future of parentage analysis

The future of parentage analysis looks at least as bright as the present. Despite the complexities we have covered in this review, the bottom line is that the techniques currently available work extremely well, and parentage analysis is completely feasible in most systems. Nevertheless, if the recent past can serve as a guide, we should expect some important developments

© 2009 Blackwell Publishing Ltd

Table 1 Recommended programs for performing the types of parentage analysis described in this article. Programs are categorized by their primary purpose, but some programs can perform other analyses as indicated by the X's. Programs that perform paternity/maternity can assign parents singly to an offspring (without necessarily considering the genotype of the other parent), whereas parent-pair allocation assigns both parents simultaneously to an offspring, while checking that those parents could indeed have produced the offspring genotype. We also provide a qualitative assessment of how well the programs handle null alleles, scoring errors and mutations. These latter recommendations should serve as a rough guide, but we encourage the reader to delve more deeply into the precise error-accommodation methods used by each program. See the Appendix for a more detailed description of each program

	Available functions								Error accommodation		
	PM	PP	PR	SR	IC*	EC†	EP‡	FP§	Nulls¶	Error/Mut**	Comments
Exclusion											
FAP 3.6		X				X			None	Moderate	Taggart (2007); http://www.aqua.stir.ac.uk/rep-gen/downloads.php
Categorical allocation											
CERVUS 3.0	X	X			X	X	X		Moderate	Good	Marshall *et al.* (1998); Kalinowski *et al.* (2007); http://www.fieldgenetics.com/
PASOS 1.0	X	X				X			None	Good	Duchesne *et al.* (2005); http://www2.bio.ulaval.ca/louisbernatchez/downloads.htm
Fractional allocation											
PATRI	X				X			X	None	None	Signorovitch & Nielsen (2002); http://people.binf.ku.dk/rasmus/webpage/patri.html
Full probability											
MASTERBAYES	X	X			X			X	Good	Good	Hadfield *et al.* (2006); http://cran.r-project.org/web/packages/MasterBayes/index.html
Parental Reconstruction											
GERUD 2.0			X			X	X		None	None	Jones (2001, 2005); http://www.bio.tamu.edu/USERS/ajones/JonesLab.htm
PARENTAGE 1.0			X		X				None	Moderate	Emery *et al.* (2001); http://www.mas.ncl.ac.uk/~nijw/#parentage
Sibship reconstruction											
COLONY 2.0	X	X	X	X	X				Good	Good	Wang (2004); http://www.zsl.org/science/research/software/
PEDIGREE 2.2			X	X					Poor	Poor	Smith *et al.* (2001); http://herbinger.biology.dal.ca:5080/Pedigree/

PM, paternity/maternity; PP, parent pair allocation; PR, parental reconstruction; SR, sibship reconstruction.
*Ability to assign statistical confidence for particular parent-offspring pairs.
†Ability to assess the expected confidence in assignments on an experiment-wide basis.
‡Ability to calculate exclusion probabilities.
§Full probability parentage analysis.
¶Null allele handling.
**Genotyping error/mutation handling.

© 2009 Blackwell Publishing Ltd

Table 2 Niche programs that may be useful in the analysis of data sets characterized by unusual features, such as haplo-diploidy or dominant markers

	Available functions								Error accommodation		
	PM	PP	PR	SR	IC*	EC†	EP‡	FP§	Nulls¶	Error/Mut**	Comments
Multigenerational											
PEDAPP 1.0	X				X			X	None	None	Muti-generational pedigree reconstruction; Almudevar (2007); http://www.urmc.rochester.edu/smd/biostat/people/faculty/almudevar.html
Sneaking Detection											
TWO-SEX PATERNITY	X				X				None	None	Assigns a percentage of a particular brood to a putative parent; Neff *et al.* (2000); http://publish.uwo.ca/~bneff/software.htm
Sex-Linked Markers											
NEWPAT 5	X				X	X			Moderate	Moderate	Exclusion with sex-linked markers; Worthington Wilmer *et al.* (1999); http://www.zoo.cam.ac.uk/zoostaff/amos/newpat.htm
Dominant Markers											
FAMOZ	X	X			X	X	X		Poor	Good	Assignment with co-dominant or dominant markers; Gerber *et al.* (2003); http://www.pierroton.inra.fr/genetics/labo/Software/Famoz/
Automated Data Handling											
PAE	X						X		None	None	Assignment that interfaces directly with fragment analysis software; Rocheta *et al.* (2007); http://www.math.ist.utl.pt/~fmd/pa/pa.zip
Closed Systems											
PAPA 2.0	X	X				X			None	Good	Parent-pair assignment for breeding experiments in closed systems; Duchesne *et al.* (2002); http://www2.bio.ulaval.ca/louisbernatchez/downloads.htm
Haplo-Diploids											
MATESOFT	X		X						None	None	Parental reconstruction in haplo-diploids; Moilanen *et al.* (2004); http://www.bi.ku.dk/staff/jspedersen/matesoft/

PM, paternity/maternity; PP, parent pair allocation; PR, parental reconstruction; SR, sibship reconstruction.
*Ability to assign statistical confidence for particular parent-offspring pairs.
†Ability to assess the expected confidence in assignments on an experiment-wide basis.
‡Ability to calculate exclusion probabilities.
§Full probability parentage analysis.
¶Null allele handling.
**Genotyping error/mutation handling.

© 2009 Blackwell Publishing Ltd

Table 3 Programs for the adventurous molecular ecologist. The functions performed by most of these programs can more easily be performed by one of the programs listed in Table 1. However, these programs may be convenient for use on some types of data sets or by individuals familiar with them

	Available functions								Error accommodation		
	PM	PP	PR	SR	IC*	EC†	EP‡	FP§	Nulls¶	Error/Mut**	Comments
Exclusion											
FAMSPHERE 0.4	X	X							None	Poor	Carvajal-Rodriguez (2007); http://webs.uvigo.es/acraaj/famsphere.htm
PROBMAX 3.0		X							Good	Moderate	Danzmann (1997); http://www.uoguelph.ca/~rdanzman/software/PROBMAX/
WHICHPARENTS 1.0	X	X							Moderate	Moderate	http://www.bml.ucdavis.edu/whichparents.html
KINSHIP 1.3.1	X			X	X				None	None	Goodnight & Queller (1999); http://www.gsoftnet.us/GSoft.html
Categorical allocation											
PARENTE	X	X			X				Poor	Good	Cercueil *et al.* (2002); http://www-leca.ujf-grenoble.fr/logiciels.htm
Sibship reconstruction											
PRT				X		X			Moderate	Poor	Almudevar & Field (1999); http://www.urmc.rochester.edu/smd/biostat/people/faculty/almudevar.html
ML-RELATE				X	X				Moderate	None	Kalinowski *et al.* (2006a,b))*et al.*; http://www.montana.edu/kalinowski/Software/MLRelate.htm
KINALYZER				X					Poor	Moderate	Ashley *et al.* (2009); http://kinalyzer.cs.uic.edu
FAMILYFINDER				X					None	Moderate	Beyer & May (2003); http://genome-lab.ucdavis.edu/People/Alumni/JenBeyer/familyFinder.html
KINGROUP 2	X			X	X				None	None	Konovalov *et al.* (2004); http://code.google.com/p/kingroup

PM, paternity/maternity; PP, parent pair allocation; PR, parental reconstruction; SR, sibship reconstruction.
*Ability to assign statistical confidence for particular parent-offspring pairs.
†Ability to assess the expected confidence in assignments on an experiment-wide basis.
‡Ability to calculate exclusion probabilities.
§Full probability parentage analysis.
¶Null allele handling.
**Genotyping error/mutation handling.

© 2009 Blackwell Publishing Ltd

over the next decade even though the conceptual core of parentage analysis will not change. Since the last review of parentage analysis methods (Jones & Ardren 2003), there have been some key advances. Most notably, full probability parentage analysis and sibship reconstruction have progressed enough that they can be seen as legitimate approaches to some types of parentage analysis problems. In addition, many of the existing techniques have made slight adjustments to various aspects of their algorithms to improve performance.

Where should we focus our effort in the coming years with respect to parentage analysis? The comparison of various parentage algorithms is starting to become a cottage industry (Slate *et al.* 2000; Oddou-Muratorio *et al.* 2003; Slavov *et al.* 2005; Herlin *et al.* 2007; Duchesne *et al.* 2008; Sefc & Koblmüller 2009), and this trend should continue because too few such studies have yet been conducted to produce clear generalities. Even now, very few of the methods have been rigorously tested on simulated data sets or in experimental populations with known patterns of parentage. In the future, comparisons of approaches should not only endeavour to establish which techniques provide the closest estimate of the true patterns of parentage but they also should focus on the estimation of variables of interest (Araki & Blouin 2005; Charmantier & Reale 2005). For example, if the goal is to characterize a sexual selection differential on a phenotypic trait, does one method produce a better, less-biased estimate than another? Similarly, if the goal is to estimate quantitative genetic parameters, is there a preferred method for parentage assignment? Perhaps the full probability parentage model is the one to be used under these circumstances, but we will not know for certain until we have more empirical and simulation data comparing methods.

Another goal in the development of parentage analysis software should be to make the software packages user friendly. Why is CERVUS (Marshall *et al.* 1998) such a popular program for parentage analysis? The answer is partly that it implements a sound approach and partly that it is so easy to use. Particularly as models become more complex, as in the case of full probability parentage analysis, user friendliness becomes a bigger and bigger issue. If the program has hidden assumptions, and the user is unaware of them because of a weak interface, the potential for misuse of the program increases dramatically. We encourage developers to create stand-alone programs rather than programs that are dependent on some other platform (such as the R statistical language) and to use standard data formats (such as the GENEPOP format; Raymond & Rousset 1995). Detailed user manuals also are essential.

Finally, we want to end by stressing again the most important factors that contribute to a successful study. The most difficult part of a parentage study is to obtain appropriate samples from the population of interest, so an investment of time and effort in the design of the study and in field work will pay dividends. If the field sampling is inadequate, no amount of molecular data or technical sophistication will make up for this shortcoming. An equally important requirement for success in parentage analysis is to have a reliable set of molecular markers. The identification of polymorphic markers with a low rate of genotyping error should be a goal of any study. With a good set of markers in hand and adequate samples from the field, most molecular ecologists should find parentage analysis to be a painless, rewarding experience.

Acknowledgements

We would like to thank JD Hadfield and two anonymous referees for helpful comments on this manuscript.

References

Almudevar A (2007) A graphical approach to relatedness inference. *Theoretical Population Biology*, **71**, 213–229.

Almudevar A, Field C (1999) Estimation of single generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**, 136–165.

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

Araki H, Blouin MS (2005) Unbiased estimation of relative reproductive success of different groups: evaluation and correction of bias caused by parentage assignment errors. *Molecular Ecology*, **14**, 4097–4109.

Ashley MV, Berger-Wolf TY, Caballero IC, Chaovalitwongse W, DasGupta B, Sheikh SI (2009) Full sibling reconstruction in wild populations from microsatellite genetic markers. In: *Computational Biology: New Research* (ed. Russe AS), pp. 231–258. Nova Publishers, Hauppauge, NY.

Ashley MV, Caballero IC, Chaovalitwongse W *et al.* (2009) KINALYZER, a computer program for reconstructing sibling groups. *Molecular Ecology Resources*, **9**, 1127–1131.

Avise JC (2004) *Molecular Markers, Natural History, and Evolution*. Sinauer, Sunderland, Mass.

Baruch E, Weller JI (2008) Estimation of the number of SNP genetic markers required for parentage verification. *Animal Genetics*, **39**, 474–479.

Beyer J, May B (2003) A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, **12**, 2243–2250.

Berger-Wolf TY, Sheikh SL, DasGupta B, Ashley MV, Caballero IC, Chaovalitwongse W, Putreva SL (2007) Reconstructing sibling relationships in wild populations. *Bioinformatics*, **23**, i49–i56.

© 2009 Blackwell Publishing Ltd

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

Brookfield JFY (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology*, **5**, 4534–4555.

Burke T, Bruford MW (1987) DNA fingerprinting in birds. *Nature*, **327**, 149–152.

Carvajal-Rodriguez A (2007) FAMSPHERE: a computer program for parental allocation from known genotypic pools. *Molecular Ecology Notes*, **7**, 213–216.

Cercueil A, Bellemain E, Manel S (2002) PARENTE: computer program for parentage analysis. *Journal of Heredity*, **93**, 458–459.

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

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

Chakraborty R, De Andrade M, Daiger SP, Budowle B (1992) Apparent heterozygote deficiencies observed in DNA typing and their implications in forensic applications. *Annals of Human Genetics*, **56**, 455–457.

Charmantier A, Reale D (2005) How do misassigned paternities affect the estimation of heritability in the wild? *Molecular Ecology*, **14**, 2839–2850.

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.

Danzmann RG (1997) PROBMAX: a computer program for assigning unknown parentage in pedigree analysis from known genotypic pools of parents and progeny. *Journal of Heredity*, **88**, 333.

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

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.

DeWoody JA, Walker D, Avise JC (2000a) Genetic parentage in large half-sib clutches: theoretical estimates and empirical appraisals. *Genetics*, **154**, 1907–1912.

DeWoody JA, DeWoody YD, Fiumera AC, Avise JC (2000b) On the number of reproductives contributing to a half-sib progeny array. *Genetical Research*, **75**, 95–105.

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

Dow BD, Ashley MV (1996) Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, *Quercus macrocarpa*. *Molecular Ecology*, **5**, 615–627.

Duchesne P, Godbout MH, Bernatchez L (2002) PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation. *Molecular Ecology Notes*, **2**, 191–193.

Duchesne P, Castric T, Bernatchez L (2005) PASOS (parental allocation of singles in open systems): a computer program for individual parental allocation with missing parents. *Molecular Ecology Notes*, **5**, 701–704.

Duchesne P, Meldgaard T, Berrebi P (2008) Parentage analysis with few contributing breeders: validation and improvement. *Journal of Heredity*, **99**, 323–334.

Eckert KA, Hile SE (2009) Every microsatellite is different: intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Molecular Carcinogenesis*, **48**, 379–388.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.

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

Emery AM, Wilson IJ, Craig S, Boyle PR, Noble LR (2001) Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Molecular Ecology*, **10**, 1265–1278.

Estoup A, Cornuet JM, Rousset F, Guyomard R (1999) Juxtaposed microsatellite systems as diagnostic markers for admixture: theoretical aspects. *Molecular Biology and Evolution*, **16**, 898–908.

Garant D, Kruuk LEB (2005) How to use molecular marker data to measure evolutionary parameters in wild populations. *Molecular Ecology*, **14**, 1843–1859.

Gerber S, Mariette S, Streiff R, Bodenes C, Kremer A (2000) Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Molecular Ecology*, **9**, 1037–1048.

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.

Glaubitz JC, Rhodes OE, DeWoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, **12**, 1039–1047.

Goodnight KF, Queller DC (1999) Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, **8**, 1231–1234.

Gowaty PA, Karlin AA (1984) Multiple maternity and paternity in single broods of apparently monogamous eastern bluebirds (*Sialia sialis*). *Behavioral Ecology and Sociobiology*, **15**, 91–94.

Hadfield JD, Richardson DS, Burke T (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology*, **15**, 3715–3730.

Haig SM (1998) Molecular contributions to conservation. *Ecology*, **79**, *413–425.*

Hardesty BD, Hubbell SP, Bermingham E (2006) Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters*, **9**, 516–525.

Herlin M, Taggart JB, McAndrew BJ, Penman DJ (2007) Parentage allocation in a complex situation: a large commercial Atlantic cod (*Gadus morhua*) mass spawning tank. *Aquaculture*, **272**, S195–S203.

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

Hubby JL, Lewontin RC (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, **54**, 577–594.

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

© 2009 Blackwell Publishing Ltd

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature*, **314**, 67–73.

Johnson PCD, Haydon DT (2006) Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics*, **175**, 827–842.

Jones AG (2001) GERUD1.0: a computer program for the reconstruction of parental genotypes from progeny arrays using multi-locus DNA data. *Molecular Ecology Notes*, **1**, 215–218.

Jones AG (2005) GERUD2.0: a computer program for the reconstruction of parental genotypes from progeny arrays with known or unknown parents. *Molecular Ecology Notes*, **5**, 708–711.

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

Jones AG, Avise JC (1997a) Microsatellite analysis of maternity and the mating system in the Gulf pipefish *Syngnathus scovelli*, a species with male pregnancy and sex-role reversal. *Molecular Ecology*, **6**, 203–213.

Jones AG, Avise JC (1997b) Polygynandry in the dusky pipefish *Syngnathus floridae* revealed by microsatellite DNA markers. *Evolution*, **51**, 1611–1622.

Jones AG, Kvarnemo C, Moore GI, Simmons LW, Avise JC (1998a) Microsatellite evidence for monogamy and sex-biased recombination in the Western Australian seahorse, *Hippocampus angustus*. *Molecular Ecology*, **7**, 1497–1505.

Jones AG, Östlund-Nilsson S, Avise JC (1998b) A microsatellite assessment of sneaked fertilizations and egg thievery in the fifteenspine stickleback. *Evolution*, **52**, 848–858.

Jones AG, Arguello JR, Arnold SJ (2002) Validation of Bateman's principles: a genetic study of mating patterns and sexual selection in newts. *Proceedings of the Royal Society of London B*, **269**, 2533–2539.

Jones B, Walsh D, Werner L, Fiumera A (2009) Using blocks of linked single nucleotide polymorphisms as highly polymorphic genetic markers for parentage analysis. *Molecular Ecology Resources*, **9**, 487–497.

Kalinowski ST, Taper ML (2006) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics*, **7**, 991–995.

Kalinowski ST, Taper ML, Creel S (2006a) Using DNA from noninvasive samples to identify individuals and census populations: an evidential approach tolerant of genotyping errors. *Conservation Genetics*, **7**, 319–329.

Kalinowski ST, Wagner AP, Taper ML (2006b) ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes*, **6**, 576–579.

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.

Koch M, Hadfield JD, Sefc KM, Sturmbauer C (2008) Pedigree reconstruction in wild fish populations. *Molecular Ecology*, **17**, 4500–4511.

Konovalov DA, Manning C, Henshaw MT (2004) KINGROUP: a program for pedigree relationship reconstruction and kin group assignment using genetic markers. *Molecular Ecology Notes*, **4**, 779–782.

Kruuk LEB, Clutton-Brock TH, Slate J, Pemberton JM, Brotherstone S, Guinness FE (2000) Heritability of fitness in a wild mammal population. *Proceedings of the National Academy of Sciences USA*, **97**, 698–703.

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

Matson SE, Camara MD, Elchert W, Banks MA (2008) P-LOCI: a computer program for choosing the most efficient set of loci for parentage assignment. *Molecular Ecology Resources*, **8**, 765–768.

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

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

Meudt HM, Clarke AC (2007) Almost Forgotten or Latest Practice? AFLP applications, analyses and advances. *Trends in Plant Science*, **12**, 106–117.

Mobley KB, Jones AG (2009) Environmental, demographic, and genetic mating system variation among five geographically distinct dusky pipefish (*Syngnathus floridae*) populations. *Molecular Ecology*, **18**, 1476–1490.

Moilanen A, Sundström L, Pedersen J (2004) MATESOFT: a program for deducing parental genotypes and estimating mating systems statistics in haplodiploid species. *Molecular Ecology Notes*, **4**, 795–797.

Morgan MT, Conner JK (2001) Using genetic markers to directly estimate male selection gradients. *Evolution*, **55**, 272–281.

Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution*, **19**, 208–216.

Myers EM, Zamudio KR (2004) Multiple paternity in an aggregate breeding amphibian: the effect of reproductive skew on estimates of male reproductive success. *Molecular Ecology*, **13**, 1951–1963.

Neff BD (2001) Genetic paternity analysis and breeding success in bluegill sunfish (*Lepomis macrochiros*). *Journal of Heredity*, **92**, 111–119.

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

Neff BD, Repka J, Gross MR (2001) A Bayesian framework for parentage analysis: the value of genetic and other biological data. *Theoretical Population Biology*, **59**, 315–331.

Neff BD, Pitcher TE, Repka J (2002) A Bayesian model for assessing the frequency of multiple mating in nature. *Journal of Heredity*, **93**, 406–414.

Neff BD, Pitcher TE, Ramnarine IW (2008) Inter-population variation in multiple paternity and reproductive skew in the guppy. *Molecular Ecology*, **17**, 2975–2984.

Nielsen R, Mattila DK, Clapham PJ, Palsbøll PJ (2001) Statistical approaches to paternity analysis in natural populations and applications to Northern Atlantic humpback whale. *Genetics*, **157**, 1673–1682.

Oddou-Muratorio S, Houot M-L, 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.

Olsen JB, Busack C, Britt J, Bentzen P (2001) The aunt and uncle effect: an empirical evaluation of the confounding influence of

© 2009 Blackwell Publishing Ltd

full sibs of parents on pedigree reconstruction. *Journal of Heredity*, **92**, 243–247.

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535–538.

Pemberton JM (2009) Wild pedigrees: the way forward. *Proceedings of the Royal Society of London B*, **275**, 613–621.

Pemberton JM, Slate J, Bancroft DR, Barrett JA (1995) Nonamplifying alleles at microsatellite loci – a caution for parentage and population studies. *Molecular Ecology*, **4**, 249–252.

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

Planes S, Jones GP, Thorrold SR (2009) Larval dispersal connects fish populations in a network of marine protected areas. *Proceedings of the National Academy of Sciences USA*, **106**, 5693–5697.

Ramakrishnan U, Storz JF, Taylor BL, Lande R (2004) Estimation of genetically effective breeding numbers using a rejection algorithm approach. *Molecular Ecology*, **13**, 3283–3292.

Raymond M, Rousset F (1995) GENEPOP (version 1.2) – Population-genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Rocheta M, Dionisio FM, Fonesca L, Pires AM (2007) Paternity analysis in Excel. *Computer Methods and Programs in Biomedicine*, **88**, 234–238.

Roeder K, Devlin B, Lindsay BG (1989) Application of maximum-likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, **45**, 363–379.

Sefc KM, Koblmüller S (2009) Assessing parent numbers from offspring genotypes: the importance of marker polymorphism. *Journal of Heredity*, **100**, 197–205.

Sezen UU, Chazdon RL, Holsinger KE (2009) Proximity is not a proxy for parentage in an animal-dispersed Neotropical canopy palm. *Proceedings of the Royal Society of London B*, **276**, 2037–2044.

Signorovitch J, Nielsen R (2002) PATRI – paternity inference using genetic data. *Bioinformatics*, **18**, 341–342.

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

Slavov GT, Howe GT, Gyaourova AV, Birkes DS, Adams WT (2005) Estimating pollen flow using SSR markers and paternity exclusion: accounting for mistyping. *Molecular Ecology*, **14**, 3109–3121.

Smith BR, Herbinger CM, Merry HR (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, **158**, 1329–1338.

Smouse PE, Meagher TR, Kobak CJ (1999) Parentage analysis in *Chamaelirium luteum* (L.) Gray (Liliaceae): why do some males have higher reproductive contributions? *Journal of Evolutionary Biology*, **12**, 1069–1077.

Taggart JB (2007) FAP: an exclusion-based parental assignment program with enhanced predictive functions. *Molecular Ecology Notes*, **7**, 412–415.

Tatarenkov A, Healy CIM, Grether GF, Avise JC (2008) Pronounced reproductive skew in a natural population of green swordtails, *Xiphophorus helleri*. *Molecular Ecology*, **20**, 4522–4534.

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, **17**, 6463–6471.

Teixeira S, Bernasconi G (2007) High prevalence of multiple paternity within fruits in natural populations of *Silene latifolia*, as revealed by microsatellite DNA analysis. *Molecular Ecology*, **16**, 4370–4379.

Thomas SC, Hill WG (2002) Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetical Research*, **79**, 227–234.

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

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

Thompson EA (1976b) A paradox of genealogical inference. *Advances in Applied Probability*, **8**, 648–650.

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

Walsh PS, Fildes NJ, Reynolds R (1996) Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus *vWA*. *Nucleic Acids Research*, **24**, 2807–2812.

Wang JL (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963–1979.

Wang JL (2007) Parentage and sibship exclusions: higher statistical power with more family members. *Heredity*, **99**, 205–217.

Westneat DF, Sherman PW (1997) Density and extra-pair fertilizations in birds: a comparative analysis. *Behavioral Ecology and Sociobiology*, **41**, 205–215.

Worthington Wilmer J, Allen PJ, Pomeroy PP, Twiss SD, Amos W (1999) Where have all the fathers gone? An extensive microsatellite analysis of paternity in the grey seal (*Halichoerus grypus*). *Molecular Ecology*, **8**, 1417–1429.

Yezerinac SM, Weatherhead PJ, Boag PT (1995) Extra-pair paternity and the opportunity for sexual selection in a socially monogamous bird (*Dendroica petechia*). *Behavioral Ecology and Sociobiology*, **37**, 179–188.

Appendix

In this appendix, we provide brief descriptions of the programs listed in Tables 1–3. The user guides and publications pertaining to each program provide more information, but this appendix should be a useful starting place. Programs are listed in alphabetical order, and we usually only discuss the most recent versions of each program.

CERVUS

CERVUS is the most popular categorical allocation program. It calculates a LOD score for each possible parent-offspring pairing and uses this value to assign parentage across a group of offspring. Simulations are used to determine critical values for Δ (the difference between the LOD scores of the two most likely candidate par-

© 2009 Blackwell Publishing Ltd

ents) that produce a desired level of confidence in assignments. The likelihood expressions incorporate a genotype replacement model, in which the probability of a replacement is based on the allele frequencies in the population (Marshall *et al.* 1998; Kalinowski *et al.* 2007), to account for genotypic mismatches in the data set brought on by mutation or experimental error. To facilitate these calculations, the user is asked to enter a reasonable per locus error rate for the data set. CERVUS has no formal framework for handling null alleles, but it does detect their presence and estimate their frequencies by examining departures from Hardy-Weinberg equilibrium, so the user can make informed decisions about whether to exclude affected loci in an analysis or not. Another feature of CERVUS allows the user to simulate inbreeding by incorporating information about relatedness among inbreeding parents and the rate of mating between relatives. CERVUS can assign one parent at a time or parent pairs. In the case of parent-pair analyses, the gender of the parents may be known or unknown. Although the parent assignment framework in CERVUS is most powerful when all candidate parents have been sampled, the program also accommodates open systems in which a proportion of the potential parents are missing from the sample. CERVUS also reports results in such a way that it can be used easily for complete exclusion as well.

COLONY

COLONY uses maximum likelihood to assign both sibship and parentage relationships. In this program, offspring are first clustered into paternal families and maternal families using a simulated annealing approach to maximize the group likelihood value. Under this method, all individuals are considered and partitioned simultaneously, resulting in higher power than is found in pairwise likelihood. Candidate parents are then assigned to the clusters at 95% confidence. If no candidate parents are available, or if no suitable candidate parents are found, the program will reconstruct parental genotypes. The straightforward user interface accommodates input of known relationships between individuals and confirms agreement between data sets and user-specified parameters. The program allows monogamy or polygamy, diploidy or haplodiploidy and codominant or dominant markers. The error model simultaneously handles two types of errors: null alleles or allelic dropout (type 1) and other stochastic error (type 2) such as mutation or random typing errors. This model assumes that error occurs independently across loci. Specified error is then incorporated in the group-likelihood calculations. Several freeware programs are available to organize the tab-delimited output into pedigree diagrams. This program works best with highly polymorphic markers, as less informative markers can result in incorrectly assigned parents.

FAP

FAP uses an exclusion approach to provide information on three different topics: (1) the predictive power of a given set of loci for assigning offspring to parental pairs, (2) actual assignment of offspring to parent pairs, and (3) identification of potentially problematic or misscored loci. The program is most suitable for the case in which all parents of the offspring are in the pool of candidate parents and does not allow null alleles to be used at any loci. The program predicts the resolving power of a given set of loci by simulating all possible parental pairings. The power of exclusion is then measured by the number of potential parental pairs that give rise to identical progeny genotypes. A subset of generated offspring genotypes is assigned to families and the percentage assigned to a single family provides a metric of the resolution power across the entire sample. Additionally, the proportion of progeny assigned without ambiguity to each family is used as a metric of the resolution power across families. Offspring are assigned to families by genotype matching; the genotypes of the actual offspring are matched to the genotypes produced in all of the pairwise combinations of parental genotypes in the previous step. There are two ways that FAP tolerates genotyping errors and mutations: allele mismatches and allele sizes. First, there is an option to allow for alleles to be mismatched between parents and offspring: the number of loci can range from zero to $n - 1$, where n is the number of loci used in the analysis. Second, to account for error in allele size measurement there is an option to tolerate a range of sizes of the alleles being matched between parents and offspring. The output of the program indicates which alleles were not matched and thus identifies the potentially problematic loci.

FAMILYFINDER

FAMILYFINDER uses graphical algorithms to assemble full sib groups based on pairwise likelihood ratios. Significance is established by comparing the pairwise likelihood ratio to simulated ratios of known relatedness. Pairs of individuals with a significant pairwise likelihood ratio are connected by a line. After all pairs of individuals are evaluated, full-sib groups can be identified as a group of individuals (each individual is a node) completely interconnected by lines. The program then uses connected component and minimum cut algorithms to search for and remove weak connections between the groups,

© 2009 Blackwell Publishing Ltd

which are assumed to be spurious. This program works best with large full sib families and many, highly polymorphic markers. FAMILYFINDER runs through UNIX/LINUX.

FAMOZ

FAMOZ seems to be the software of choice for categorical allocation involving data sets with dominant marker loci. FAMOZ can also include co-dominant and cytoplasmic markers in a single analysis, making it a good choice for studies involving multiple types of molecular data. Confidence is assessed by running simulations similar to those found in CERVUS, except that FAMOZ uses the raw LOD score rather than Δ; and errors also are handled using methods similar to those in CERVUS. Finally, FAMOZ includes a subroutine to calculate gene flow into the study population based on the results of the parentage assignment. FAMOZ requires the ToolCommandLanguage/Toolkit (TclTk) platform (Gerber *et al.* 2003), so it is slightly more difficult to run than some of the other parentage analysis tools.

FAMSPHERE

FAMSPHERE excludes parent pairs (or single parents if one parent is known *a priori*) for individual offspring on the basis of genotypic data. The user can define a positive integer as the radius R, which is assumed to be proportional to mutation/genotyping error. In the case of multiple nonexcluded parent pairs or single parents, FAMSPHERE utilises a distance-based graphical method to allocate the parents.

GERUD

GERUD is a program that reconstructs parental genotypes from progeny arrays. The progeny array must contain only half- and full-sibs, with or without a known parental genotype. If neither parental genotype is known, GERUD will find all genotypes that could represent a shared parent among all progeny. The program then uses an exhaustive algorithm that tests all possible parental genotypes against the progeny array to find the minimum number of genotypes necessary to explain the array. This approach is guaranteed to find the minimum number of parents, and it takes into account associations of alleles across loci. In addition to determining the minimum number of parents, GERUD reports their genotypes. The program also includes a simulation routine, which uses allele frequencies to determine the probability that the minimum number of parents reported by GERUD will represent the true number of parents and the proportion of genotypes expected to be correctly reconstructed.

KINALYZER

KINALYZER uses a combinatorial optimization approach to create full sibling groups. The objective of the combinatorial optimization approach is to find the solution in which the fewest number of full-sib groups are produced. The software is web-based and has an easy user interface. Data can be analysed using either the Minimum Two Allele Set Cover approach or the Consensus algorithm. The Minimum Two Allele Set Cover algorithm groups individuals as siblings if they follow the rules of Mendelian inheritance at all loci. The Consensus algorithm removes loci one at a time from the data to find partial solutions, and then reincorporates the removed loci. The consensus result is calculated by finding the kin groups common to all the partial solutions and iteratively incorporating additional groups and individuals. This algorithm is more tolerant to low allelic variation and error. The program does not group half-sibs.

KINGROUP

KINGROUP is a JAVA based program that uses maximum pairwise likelihood to group relatives. The program provides several algorithms (descending ratio, exhaustive descent, Simpson, and others) to group related individuals. Candidate partitions are built by adding single individuals to existing groups. In the descending ratio algorithm, those partitions with the highest likelihood value are retained for the next iteration. In the exhaustive descent algorithm, all candidate partitions are retained for future iterations. KINGROUP can also accommodate haplo-diploid organisms.

KINSHIP

KINSHIP is a useful program for calculating pair-wise relatedness values among all pairs of individuals in a sample. It also can be used to identify potential parent-offspring pairs. KINSHIP provides a confidence score for individual assignments, but does not adequately address statistical confidence, severely impairing its usefulness for parentage assignment.

MASTERBAYES

MASTERBAYES is a package implemented in the R statistical programming language for the evaluation of full probability parentage models. The approach is to specify a model, which can include multiple parts and parameters in addition to the patterns of parentage. The goal is to determine values of the parameters, including the parent-offspring relationships that maximize the overall posterior probability. MASTERBAYES uses a Markov Chain

© 2009 Blackwell Publishing Ltd

Monte Carlo approach to find the best solution. In principle, nearly any full probability model could be implemented by using MASTERBAYES, so the approach is quite flexible. For example, MASTERBAYES can include a robust model for handling genotyping errors, mutations, and null alleles, and it can also simultaneously estimate the proportion of candidate parents in a way similar to the approach used by PATRI. However, familiarity with R is a necessity for using MASTERBAYES effectively, so it may not experience as wide of usage as a stand-alone program like CERVUS.

MATESOFT

MATESOFT is designed for haplo-diploid systems. The program has four capabilities: (1) Constructing maternal genotypes given the genotypes of a group of sisters, (2) constructing paternal genotypes from a queen and her offspring (possibly using the reconstructed maternal genotype from step 1), (3) calculation of mating frequency statistics (i.e., paternity skew, contribution of each father for multiple-sire families, proportion of doubly mated females in the population, and effective mate number for females), and (4) calculation of mating frequency statistics for an inseminated female with stored sperm. The program is fairly straightforward and easy to use, though limited in the ability to accommodate genotyping errors, null alleles and mutations.

ML-RELATE

ML-RELATE uses maximum likelihood estimates of relatedness to classify pairs of related individuals as unrelated, half-sibs, full-sibs, or parent-offspring relationships. The output lists all levels of relatedness that could apply to a given pair of individuals. The significance of a given level of relatedness between two individuals can be tested by the user on a pair-by-pair basis. The input for this program requires both the genotypes of sampled individuals and population level allele frequencies. Null alleles can be accommodated; the user can either specify which loci contain null alleles or test for null alleles using a Hardy-Weinberg test for heterozygote deficiency. When null alleles are specified, the program uses a maximum likelihood estimate to incorporate the frequency of null alleles in the calculations.

NEWPAT

NEWPAT is an exclusion-based approach that finds a niche by allowing sex-linked markers in the analysis. NEWPAT accommodates errors by allowing the user to specify the number of mismatches necessary for exclusion. Its method for handling confidence in individual assignments has yet to be validated, but it does provide a simulation-based approach for assessing experiment-wide confidence in exclusion.

PAE

PATERNITY ANALYSIS IN EXCEL (PAE) is a categorical allocation program that converts raw fragment analysis data to the proper format and assigns paternity (maternity) automatically, all in Microsoft Excel spreadsheets. PAE uses the same likelihood expressions as CERVUS 2.0, which should be upgraded to use the more up-to-date model implemented in CERVUS 3.0. The PAE program requires very specific sample nomenclature, so the use of this program must be anticipated during the early stages of data collection.

PAPA

PAPA is a categorical allocation program for inferring parent pairs in closed systems. It boasts a streamlined, user-friendly interface in addition to an excellent manual and help files. This software can also help optimize the number of loci required for desired assignment confidence via the 'pre-parental' simulation module. PAPA can incorporate prior information about known mating patterns using the 'crossing/matching plans' option. The gender of candidate parents may be known or unknown for parent pair assignments.

PARENTAGE

PARENTAGE is a Bayesian parental reconstruction program. It applies to full- or half-sib families and indirectly calculates the posterior probability of parentage in the array by fitting probability models to the data. The program uses a Markov Chain Monte Carlo approach to explore the sample space, which is constrained to include the parental genotypes possible given the observed offspring genotypes. The program uses an infinite alleles' model to incorporate mutation and mis-scored alleles. This program can be seen as the maximum-likelihood equivalent of GERUD, so it may provide better estimates of parental reconstruction when markers are less informative. However, PARENTAGE has the potential to overestimate the number of parents contributing to a progeny array, while GERUD does not. As markers become sufficiently informative, both approaches should converge on the same answer.

PARENTE

PARENTE performs categorical assignment for single parents or for parent pairs by using co-dominant marker

© 2009 Blackwell Publishing Ltd

data. PARENTE was one of the first programs to perform parent-pair assignment, but CERVUS has now implemented such a procedure. PARENTE uses posterior probabilities for assignment, and it allows for errors and missing parents in the pool of candidates. PARENTE also can take into account the ages of the candidate parents when performing assignments.

PASOS

PASOS employs a likelihood framework to allocate parent pairs among a group of offspring, and it accommodates open systems. Simulations incorporating user-defined parameter values are conducted before parent-pair allocation to establish experiment-wide assignment confidence and to optimize the subset of loci ultimately used for allocation. Error modelling in the simulations is achieved first through a user-defined maximum offset tolerance (MOT), which dictates the maximum number of 'mutational' steps (in either direction) allowed for any one allele during the simulations (Duchesne *et al.* 2005). A distribution of the error rate across the different stepwise deviations from the 'original' allele is specified by the user, and is assumed to be a realistic model for mismatches arising as a consequence of mutation and genotyping errors. The error model '0.002, 0.008, 0.98, 0.008, 0.002', for example, assumes a MOT of 2, a unidirectional one-step error rate of 0.008, and a unidirectional two-step error rate of 0.002. After simulations are complete, the likelihood of each real parent pair/offspring trio is calculated assuming a fixed error model, in which a general error rate of 0.02 is distributed uniformly across all alleles. For each offspring the parent pair with the highest likelihood is allocated, provided there are no allelic mismatches. If one or both parents in the most likely parent pair mismatch an offspring allele, the candidate parent(s) must then be allocated or rejected. If the mismatch offset is less than or equal to the MOT, the parent in question is allocated to the offspring. If the mismatch offset is greater than the user-defined MOT, however, the offspring's parent(s) will be deemed 'uncollected.' This process results in an estimate of the proportion of unsampled adults for a given group of offspring. Much like CERVUS, PASOS is capable of conducting parent pair allocation when the sex of candidate parents is either known or unknown.

PATRI

PATRI is a fractional allocation program, but it also possesses some features of full probability modelling. PATRI uses a Bayesian framework to calculate posterior probabilities of assignment of fathers to offspring, given a known mother. These posterior probabilities can be used for fractional allocation. In principle, they can also be used for categorical allocation, and the magnitude of the posterior probability provides a guide to confidence in assignment. PATRI also possesses the ability to estimate the proportion of candidate parents sampled in an open system and the difference in average reproductive success between two classes of individuals (e.g., dominant versus subordinate).

PEDAPP

PEDAPP considers the pedigree to be the primary object of inference, permitting calculation of the probability that individual *i* is the parent of individual *j*, via the Bayesian posterior probability. Multiple generations may be considered by coding the ages of sampled individuals accordingly. The sex of candidate parents in the sample may also be included, if known. The user interface is well-designed and straightforward. No statistical summary of the sampled pedigrees is provided, so the user must calculate posterior probabilities independently using the output.

PEDIGREE

PEDIGREE is a web-based sibship reconstruction program that uses pairwise likelihood to build partitions, or groups of related individuals, and evaluates the partitions based on a weighted score. A Markov Chain Monte Carlo approach is used to find the partition with the highest score. Individuals can be partitioned into both full sibling and half sibling groups. Full-sib groups can be nested within half-sib groups, but this arrangement only accommodates multiple mating in one parent. Nested data are presented in HTML formatted tables only, which can make exporting these results somewhat difficult. The significance of any given group can be estimated by comparing the cohesion and repulsion scores of several groups. These scores are calculated by taking the average log likelihood ratio of sibship among all individuals in a group. This measure is sensitive to group size: small groups could have inflated scores of cohesion by chance, indicating that the individuals within the group are more closely related than they actually are. For this reason, cohesion and repulsion scores can only be compared between groups of the same size. The current version of PEDIGREE does not incorporate error or significance testing. The program can also reconstruct parental genotypes. This feature lists all potential parental genotypes that could generate the offspring of a single group at each locus. Chi-square values are also provided to indicate a measure of fit for each parental genotype, but no probability values are provided.

© 2009 Blackwell Publishing Ltd

PROBMAX

PROBMAX is an exclusion-based program that can take advantage of dominant or co-dominant marker data. It is especially suited to cases in which both parents of each offspring are likely to be in the pool of candidate parents. Under these circumstances, it indicates the parent pairs compatible with each offspring genotype. PROBMAX allows a user-specified number of mismatches, including stepwise mutations, to accommodate scoring errors and mutations.

PRT

PRT uses a combination of group likelihood and exclusion principles to construct full-sib groups. The program first assembles maximal feasible sibling groups (MSG), which contain all individuals in the sample that could be produced by two hypothetical parents. This method uses the principle of exclusion to rule out sibling groups that have incompatible sibling genotypes. This is similar to the way exclusion is used to rule out genotypically incompatible parents in parentage assignment programs. In the MSG step, individuals can belong to more than one group. These groups are then assigned a likelihood-based score, which is used to construct a partition of sibling groups where each individual is part of only one group. The partition with the highest likelihood is found using a Markov Chain Monte Carlo approach.

TWO-SEX PATERNITY

TWO-SEX PATERNITY deals with the case in which a group of related progeny can be collected with a single putative parent. The program then estimates the proportion of offspring parented by the putative parent and provides confidence limits. This program is especially useful for species with nest-holding males in which cuckoldry is suspected.

WHICHPARENTS

WHICHPARENTS is an exclusion-based program that returns potential parents and parent pairs for a given offspring genotype. The program accommodates genotyping error and mutations by tolerating mismatched alleles at a level specified by the user. The program does not accommodate null alleles in a meaningful way. Hence, WHICHPARENTS is a bare bones exclusion program.

© 2009 Blackwell Publishing Ltd