MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 1158–1170 doi: 10.1111/mec.12138

Relative accuracy of three common methods of parentage analysis in natural populations

HUGO B. HARRISON,*†‡[1] PABLO SAENZ-AGUDELO,§[1] SERGE PLANES,‡¶ GEOFFREY P. JONES*† and MICHAEL L. BERUMEN§**

*School of Marine and Tropical Biology, James Cook University, Townsville, Qld 4811, Australia, †Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Qld 4811, Australia, ‡USR 3278 CRIOBE CNRS-EPHE, CBETM de l'Université de Perpignan, 66860 Perpignan Cedex, France, §Red Sea Research Center, King Abdullah University of Science and Technology, 23955-6900 Thuwal, Kingdom of Saudi Arabia, ¶Laboratoire d'Excellence "CORAIL", BP 1013 Papetoai, 98729 Moorea, French Polynesia, **Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

Abstract

Parentage studies and family reconstructions have become increasingly popular for investigating a range of evolutionary, ecological and behavioural processes in natural populations. However, a number of different assignment methods have emerged in common use and the accuracy of each may differ in relation to the number of loci examined, allelic diversity, incomplete sampling of all candidate parents and the presence of genotyping errors. Here, we examine how these factors affect the accuracy of three popular parentage inference methods (COLONY, FAMOZ and an exclusion-Bayes' theorem approach by Christie (Molecular Ecology Resources, 2010a, 10, 115) to resolve true parent–offspring pairs using simulated data. Our findings demonstrate that accuracy increases with the number and diversity of loci. These were clearly the most important factors in obtaining accurate assignments explaining 75–90% of variance in overall accuracy across 60 simulated scenarios. Furthermore, the proportion of candidate parents sampled had a small but significant impact on the susceptibility of each method to either false-positive or false-negative assignments. Within the range of values simulated, COLONY outperformed FaMoz, which outperformed the exclusion-Bayes' theorem method. However, with 20 or more highly polymorphic loci, all methods could be applied with confidence. Our results show that for parentage inference in natural populations, careful consideration of the number and quality of markers will increase the accuracy of assignments and mitigate the effects of incomplete sampling of parental populations.

Keywords: accuracy, COLONY, FaMoz, microsatellite, parentage analysis

Received 23 July 2012; revision received 10 October 2012; accepted 16 October 2012

Introduction

Our ability to infer genealogical relationships amongst individuals has become an effective approach to investigate a wide variety of evolutionary, ecological and behavioural questions. Pedigrees, often based on a combination of observation and molecular data, have given us invaluable insights into mating systems, revealing the prevalence of extra-pair paternities and cooperative breeding in the wild (e.g. Richardson *et al.* 2001; Magrath *et al.* 2009), mating behaviour and reproductive success (Araki *et al.* 2007; Rodriguez-Munoz *et al.* 2010; Ford *et al.* 2011; Kanno *et al.* 2011; Beldade *et al.* 2012) and kin association (e.g. Reeve *et al.* 1990; Buston *et al.* 2007; Piyapong *et al.* 2010) in diverse animal groups. Parentage studies and sibship reconstructions have also become increasingly popular approaches to estimate population parameters such as self-recruitment (Jones

Correspondence: Hugo B. Harrison, Fax: +61 7 47 25 1570; E-mail: hugo.harrison@my.jcu.edu.au

[1]These authors contributed equally.

et al. 2005; Saenz-Agudelo *et al.* 2009, 2012), fine-scale population structure (e.g. Nussey *et al.* 2005; Slavov *et al.* 2010) and population connectivity in the form of migration (Nathan *et al.* 2003; Harrison *et al.* 2010) or dispersal (e.g. Garcia *et al.* 2005, 2007; Jordano *et al.* 2007; Planes *et al.* 2009; Christie *et al.* 2010b; Saenz-Agudelo *et al.* 2011, 2012; Berumen *et al.* 2012; Harrison *et al.* 2012a). Parentage studies have revealed new aspects of inbreeding and trait heritability (Ritland 2000; Garant & Kruuk 2005; Pemberton 2008; Nielsen *et al.* 2012), genetic adaptation of wild species to captivity (Christie *et al.* 2012) and assisted in the restoration of captive and endangered populations (Keller & Waller 2002; Herbinger *et al.* 2006). Individual-level analyses can resolve family relationships in a wide range of taxa where this information has proven difficult to obtain from direct observations.

Recent technical advances in both the isolation of molecular markers, notably microsatellites or SNPs, and the high-throughput screening of multilocus genotypes are likely to make parentage studies more widely accessible to ecologists studying wild populations (Selkoe & Toonen 2006; Gardner *et al.* 2011; Guichoux *et al.* 2011). However, despite a proliferation of statistical approaches to infer pedigree structure or kinship relationships amongst pairs of individuals in natural populations (reviewed in Blouin 2003; Jones & Ardren 2003; Jones *et al.* 2010), parentage analysis remains a relatively new procedure. A number of different approaches are currently being used, but the factors affecting the relative accuracy of the different approaches have received little attention.

The methods used to identify parent–offspring relationships can be broadly divided into four categories: strict exclusion, categorical assignment, fractional assignment and pedigree reconstruction (Jones *et al.* 2010). Amongst these, the most commonly used methods are strict exclusion and categorical assignment, whereby the genotype of each offspring is compared to the genotype of all candidate parents. For strict exclusion methods, any parent failing to share at least one allele at a given locus is excluded. If more than one parent cannot be excluded, categorical assignments measure the likelihood of each putative parent–offspring pair of being true given their respective multilocus genotype and the observed allelic frequencies in the population (Marshall *et al.* 1998; Nielsen *et al.* 2001; Gerber *et al.* 2003; Kalinowski *et al.* 2007). Categorical assignment approaches offer several advantages over strict exclusion methods (Danzmann 1997; Goodnight & Queller 1999) as they can more easily accommodate scoring errors, missing data or null alleles that commonly occur in microsatellite data sets (Pemberton *et al.* 1995; Dakin & Avise 2004; Pompanon *et al.* 2005; Wang 2010). However, in the right circumstances, strict exclusion can be a powerful approach and could prove useful to detect parent–offspring pairs in large open populations (Christie 2010a). Recently, full-probability approaches for parental or sibship reconstructions have also become more accessible and widely applied. Rather than simply evaluating pairwise relationships, individuals are clustered into family groups and the likelihood of different clusters is evaluated to identify the most parsimonious configuration (Almudevar & Field 1999; Thomas & Hill 2002; Wang 2004; Hadfield *et al.* 2006; Wang & Santure 2009; Jones & Wang 2010a; Almudevar & Anderson 2011). In turn, accounting for the presence of family groups provides valuable information that significantly enhances the accuracy of assignments (Wang 2007; Walling *et al.* 2010).

All the above methods are subject to incorrect assignments that may be affected by the number and allelic diversity of loci examined (Bernatchez & Duchesne 2000; Nielsen *et al.* 2001), the proportion of the population sampled (Oddou-Muratorio *et al.* 2003; Koch *et al.* 2008), genotyping errors, mutations, allelic dropouts and miscalling (Bernatchez & Duchesne 2000; Hoffman & Amos 2005). However, having only a limited number of genetic markers and incomplete sampling of all candidate parents is thought to have the largest effects on the accuracy of assignments (Marshall *et al.* 1998; Nielsen *et al.* 2001; Wilson & Ferguson 2002; Oddou-Muratorio *et al.* 2003; Jones *et al.* 2010). Some likelihood-based approaches such as CERVUS (Marshall *et al.* 1998; Kalinowski *et al.* 2007) and full-likelihood methods such as COLONY (Wang 2004; Jones & Wang 2010a) account for incomplete sampling by defining a priori the probability that the true parent is present in the sample. This probability can be estimated from the proportion of putative parents sampled from the entire parental population, which requires prior knowledge, or approximation, of the size of the population. Whilst COLONY is robust to uncertainty in this sampling rate (Wang & Santure 2009; Jones & Wang 2010b), misspecification of this parameter in CERVUS can have significant impact on assignments made (Nielsen *et al.* 2001; Hadfield *et al.* 2006; Koch *et al.* 2008).

Other approaches have been developed to infer parentage without prior knowledge of population size or the proportion of candidate parents in the sample. Such methods have been favoured to assess population connectivity in large populations (mostly plants and marine fish) where accurate estimates of the breeding population size are often difficult to obtain. For instance, the pairwise-likelihood method implemented in FAMOZ (Gerber *et al.* 2003) estimates the likelihood ratios [log of the odds ratio (LOD) scores] of putative parent–offspring pairs being true and determines critical

thresholds to accept or reject assignments by simulating true and false parent–offspring pairs. The calculation of LOD scores is based on the same approach as CERVUS (Meagher & Thompson 1986; Marshall *et al.* 1998; Gerber *et al.* 2000); however, FAMOZ does not require a priori information of the proportion of candidate parents in the sample in order to determine critical LOD thresholds. The exclusion-Bayes' theorem approach by Christie (2010a) is another method that follows in this category. It consists of calculating the probability of false parent–offspring pairs in a data set to determine whether all putative parent–offspring pairs can be accepted with strict exclusion. In situations where the data set lacks sufficient power, Bayes' theorem is used to determine the probability of putative parent–offspring pairs being false given the frequencies of shared alleles. This approach was designed for situations where only a small fraction of all candidate parents can be sampled and does not require a priori information of the proportion of candidate parents in the sampled population or other demographic parameters (Christie 2010a). Whilst the effects of the misspecification of the proportion of sampled candidate parents in CERVUS have been evaluated and discussed elsewhere (Nielsen *et al.* 2001; Hadfield *et al.* 2006; Koch *et al.* 2008), it is unclear how the absence of this parameter may affect the performance of exclusion and categorical assignment approaches, such as those implemented in FAMOZ and exclusion-based approaches, especially under different sampling rates.

The aim of this study was to assess the accuracy of three popular methods of parentage analysis and investigate their susceptibility to error under 60 different scenarios that incrementally simulate the number of loci, allelic diversity, adult sample size and genotyping error. Simulated offspring were assigned to single parents using the exclusion-Bayes' theorem approach developed by Christie (2010a) (hereafter referred to as 'the Christie method'), the pairwise-likelihood method implemented in FAMOZ (Gerber *et al.* 2003) and the full-probability approach implemented in version 2.0 of COLONY (Wang 2004). Putative parent–offspring pairs were validated against known true parents, and assignment errors were classified as described in **Box 1**. We then examined how the number of assignment errors was correlated with the number of loci, allelic diversity and proportion of adults sampled from the population.

Materials and methods

Simulated data sets

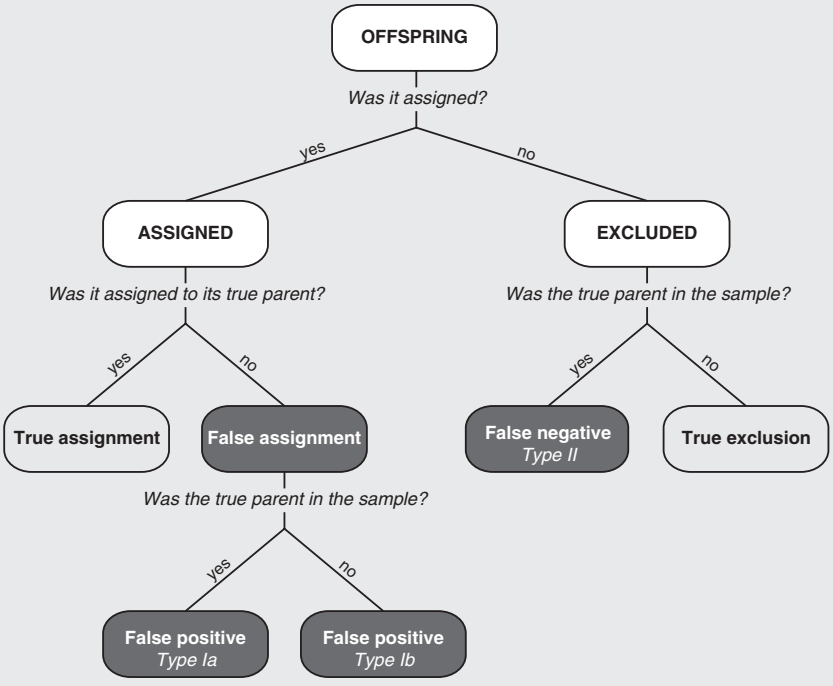
Two parental data sets, of different population sizes, were generated in EASYPOP (Balloux 2001) to achieve different levels of allelic diversity whilst maintaining all remaining simulation parameters constant. Whilst the difference in population size between both data sets has little relevance to the accuracy of assignments, this procedure allowed us to explore the effects of allelic diversity on assignments. The two parental data sets were based on a finite island model with five subpopulations, each of constant size and equal sex ratio. The first data set consisted of 500 reproductive individuals with 100 individuals per subpopulation. The second data set consisted of 1000 reproductive individuals with 200 individuals per subpopulation. These will subsequently be referred to as the N500 (low-diversity) and N1000 (high-diversity) populations, respectively. For both data sets, random mating was simulated to produce diploid genotypes at 20 independent loci for 5000 generations to approximate mutation–drift equilibrium (Waples & Gaggiotti 2006). Migration between subpopulations occurred with a probability of 0.15 to simulate high gene flow and demographic connectivity amongst subpopulations. This is equivalent to 15 and 30 migrants per generation for the N500 and N1000 populations, respectively. All loci had the same mutation dynamics, which occurred according to the K-allele model (each mutation equally likely to occur at any of K possible sites). Mutation rate ($\mu = 1 \times 10^{-4}$) and number of allelic states (20 possible allelic states) were considered to represent highly polymorphic markers, such as microsatellites, within the ranges published in eukaryotic genomes (Buschiazzo & Gemmell 2006). Our simulated data sets represented an assorted array of loci akin to most microsatellite data sets. Individual locus characteristics for each simulated data set were calculated in GENALEX v6.4 (Peakall & Smouse 2006). The N1000 population represented a more diverse and therefore informative data set with an average of 14.9 (11–18) alleles per locus and average observed heterozygosity of 0.769 ± 0.070 SD (0.650–0.877) per locus (Table S1, Supporting information). In comparison, the N500 populations had lower allelic diversity with an average of 10.7 alleles per loci (7–14) and an observed heterozygosity of 0.655 ± 0.144 SD (0.396–0.874; Table S2, Supporting information). The probability of exclusion of each locus and the cumulative probability of exclusion of each data set were calculated according to Jamieson & Taylor (1997) as the probability of excluding a single parent (Tables S1 and S2, Supporting information).

For each of the two parental data sets, 1000 offspring genotypes were generated using the software package P-LOCI (Matson *et al.* 2008). Adults were paired randomly within each subpopulation, and four offspring were generated for each adult pair under a monogamous mating system. This resulted in 250 adult pairs,

which was necessary to keep offspring sample size equal between data sets and reduce computation time of parentage analyses. Offspring were generated following Mendelian inheritance with 0.1% and 1% genotyping error, which are typical of microsatellite loci (Pompanon *et al.* 2005). Each parental population was randomly sampled into samples representing 20, 40, 60, 80 and 100 percentage of the parental population, and the resulting data sets were further subset taking the first 10, 15 and 20 loci, totaling 60 independent data sets. All data sets were deposited in the Dryad Digital Repository (Harrison *et al.* 2012b). Each of these data sets was then analysed using the following three freely available software packages to identify parent–offspring pairs.

Box 1 Measuring the accuracy of assignments and error types

Whilst the means to identify parent–offspring, half- and full-sib relationship vary widely, the objective of all parentage studies is to accurately identify these relationships. Incomplete sampling of all potential parents and insufficient number of loci are the most common cause of incorrect assignments. Here, we synthesize the decision process and errors that lead to correct or incorrect assignments and how these affect the accuracy of assignment.



Decisions and error types

There are only two correct decisions with regard to single parent assignments, assigning the true parent when it is present in the sample (true assignment) and assigning no parent when the true parent is not in the sample (true exclusion). Assignment errors can be either false positive (falsely assigning an individual to a parent that is not its true parent) or false negative (falsely excluding a true parent). These are commonly referred to as type I (false-positive) and type II (false-negative) errors, respectively, and can be estimated from simulations (e.g. FAMOZ). False positives fall into two categories, falsely assigning to a parent when the true parent is in the sample or when the true parent is not in the sample. To distinguish these from error estimates, we refer to these here as type Ia and type Ib errors, respectively. We refer to false negatives, falsely excluding a parent when it was in the sample, as a type II error. These errors cannot be calculated in real data sets unless the full pedigree is available.

Measuring the accuracy of assignments

In simulated data sets, we can identify each error type (type Ia, type Ib and type II) and measure the accuracy of each method to identify true parents (accuracy of assignments), exclude the false parents (accuracy of exclusion) or calculate the overall accuracy of assignments. We considered type Ia and type II errors to be false assignments and type Ib errors to be false exclusions. The accuracy is calculated as the sum of these errors over all possible assignments or exclusions. The overall accuracy is the sum of all errors over the total number of possible assignments.

Exclusion-Bayes' Theorem—Christie method

The method described by Christie (2010a) is an unbiased exclusion probability designed to identify true parent–offspring pairs in large populations where the proportion of sampled parents is low. For each data set, we calculated the probability of observing shared alleles between unrelated individuals using 1000 simulations, then calculated the probability of each putative parent–offspring pair being false given the frequency of shared alleles. This method does not explicitly account for genotyping error or marker-specific error rates, but allows for mismatches between parent–offspring pairs. Assignments are made to single parents only, and parent pairs in the sample are not considered. When assigned, each parent–offspring pair is given a probability and several adults may be assigned to the same offspring. When two or more putative parents were assigned to the same offspring, only the parent with the highest probability of assignment was kept for further analyses. This method was implemented in R v2.14.0 (R Development Core Team 2011).

Pairwise-Likelihood—FAMOZ

The software program FAMOZ (Gerber *et al.* 2003) allows for the categorical allocation of parent–offspring pairs based on a maximum-likelihood approach. The program computes LOD scores for assigning individuals to candidate parents based on the observed allelic frequencies at each locus. We allowed for genotyping errors by introducing an error rate of 0.01% in the LOD score calculation, which produces the lowest type I and II error rates (Gerber *et al.* 2000; Morissey and Wilson 2005; Saenz-Agudelo *et al.* 2011; Harrison *et al.* 2012a). For each data set, 10 000 parent–offspring pairs were simulated based on the observed allelic frequencies at each locus and 10 000 parent–offspring pairs were generated from the putative parental genotypes. The frequency distributions of the two simulations were compared, and the intersection was defined as the minimum threshold to accept a given parent–offspring pair or parent-pair trio. When two or more putative parents were assigned to the same offspring, only the parent with the highest LOD score was retained. When two parents were assigned as a parent pair, both were retained for further analyses.

Full-Likelihood—COLONY

The software program COLONY (Wang 2004; Jones & Wang 2010a) implements a full-likelihood approach to parentage analysis. In our analyses, we considered both parent–offspring relationships and sibship amongst offspring samples. Adult samples were separated by sex, and we assumed a polygamous mating system for diploid organisms. The prior probability that the true parent was present in the sample was considered in the assignment of parent–offspring pairs in accordance with the proportion of candidate parents included in the simulated data sets. Allelic frequencies were determined from the sample data set, but did not take into account the relationship between individuals or inbreeding. All results were based on a single short run with high precision to maximize the accuracy of assignment whilst reducing the length of individual runs. This approach accounts for genotyping error at each locus of each sampled individual when estimating the likelihood of a particular family cluster, and simulated error rates were taken into account in each analysis. Only the parent or parent pair with the highest likelihood is assigned, and all assigned parents were retained for further analysis.

Assignment errors

For each offspring, the assigned parent or parent pairs were compared to the known true parents. When an offspring was assigned to a parent that was not its true parent or not assigned (excluded), we determined whether the true parent was in the sample and identified it as either false-positive (type Ia or type Ib) or false-negative (type II) errors (**Box 1**). We then used a generalized linear model (GLM) framework to quantify the effect of allelic diversity, number of loci, percentage of sampled parents, genotyping error and their possible interaction on the proportion of correct assignments of each method. Because the response variable was a proportion, GLMs were fitted using a logit link function (as fitted values are bound between 0 and 1) and quasi-binomial errors (to account for non-normally distributed errors, nonconstant variance and overdispersion; Crawley 2007). For each method, we first fitted a maximal model (four parameters and their interactions) and then removed nonsignificant terms until a minimal adequate model was reached (Crawley 2007). Processing of all software outputs and all model fitting was performed in R with scripts deposited in the Dryad Digital Repository (Harrison *et al.* 2012b).

Results

Relative accuracy of the three methods

Given high diversity (N1000) and sufficient number of loci, all methods tested identified parent–offspring pairs with over 90% accuracy, regardless of the proportion of the population sampled or the presence of genotyping

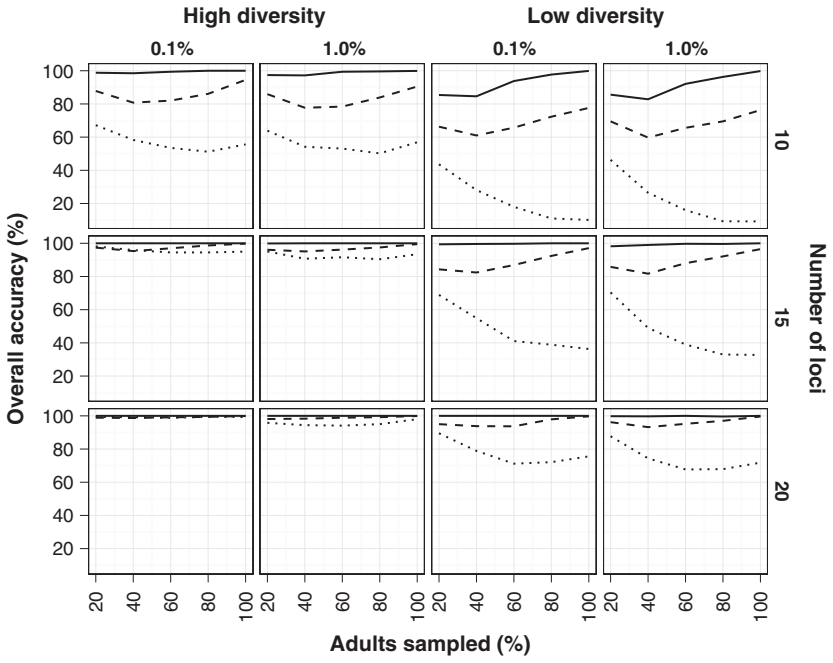


Fig. 1 Proportion of accurate assignments of three approaches to parentage analyses. Each method was tested on high- and low-diversity simulated microsatellite data sets with high (1%) and low (0.1%) levels of genotyping error for varying levels of number loci and proportion of candidate parents sampled. Continuous lines correspond to the results from the full-likelihood method implemented in COLONY v2.0 (Wang 2004), dashed lines are the results from the pairwise-likelihood implemented in FAMOZ (Gerber *et al*. 2003) and dotted lines from the Christie method (Christie 2010a).

error (Fig. 1 and Table S3, Supporting information). However, the performance of each method varied, with accuracy affected by the number and allelic diversity of loci and the proportion of sampled parents. Overall, the full-likelihood method implemented in COLONY (Wang 2004; Jones & Wang 2010a) outperformed the other two methods with a mean (±SD) accuracy across all scenarios of 98.4 ± 4.0% compared to 89.0 ± 11.3% for the pairwise-likelihood method (Gerber *et al*. 2003) and 65.3 ± 28.3% for the Christie method (Christie 2010a). For each method, the number of loci and differences in allelic diversity between the low- and high-diversity populations had the largest effect on the overall accuracy of assignments. For most scenarios we analysed, COLONY performed best, FAMOZ was intermediate and the Christie method was least accurate, with the disparity between methods increasing with increasing proportions of the population sampled (Fig. 1).

The number of loci was always the most important single factor in determining the accuracy of assignments for all three methods investigated (Fig. 1 and Tables S3–S5, Supporting information). Across all scenarios, the Christie method was the most affected with an overall reduction of 47% in overall accuracy when reducing the number of loci from 20 to 10. However, this only determined ~45% of the variance in overall accuracy (GLM: $F_{1,57} = 531.7$, $P < 0.001$), suggesting other factors are influencing the discrimination of true parent–offspring pairs. In contrast, the accuracy of the pairwise-likelihood method was reduced by only 21% overall and represented ~66% of the variance in overall accuracy ($F_{1,57} = 548.5$, $P < 0.001$). For the full-likelihood method, overall accuracy was only reduced by 4% between 20 and 10 loci, which represented ~52% of variance ($F_{1,57} = 286.1$, $P < 0.001$).

Differences in allelic diversity between the two simulated populations further accentuated the effect of the number of loci on the overall accuracy of assignment, with a significant interaction between these two factors (Table S5, Supporting information). The performance of both the Christie method and the pairwise-likelihood methods was most severely affected by their combined effect (as the sum of variances explained by each variable and their interaction), explaining a total of 90% and ~87% of variance in overall accuracy, respectively. For both methods, this included a low but significant interaction between the number of loci and allelic diversity (1.8%: $F_{1,53} = 21.9$, $P < 0.01$ and 0.8%: $F_{1,54} = 6.3$, $P < 0.05$, respectively). In contrast, these two factors explained ~75% of the overall variance in accuracy of the full-likelihood method, and there was no significant interaction between the two on the overall accuracy of assignment. Although the accuracy of the full-likelihood method was high overall, it is likely that the presence of full-sibs in our simulated data increased the accuracy of assignment for this method.

The proportion of sampled parents had a small but significant effect on the accuracy of all methods, which were only exacerbated by variation in the number and allelic diversity of loci. Variation in the proportion of sampled parents explained only ~6% of variance in overall accuracy of both the Christie method and the pairwise-likelihood

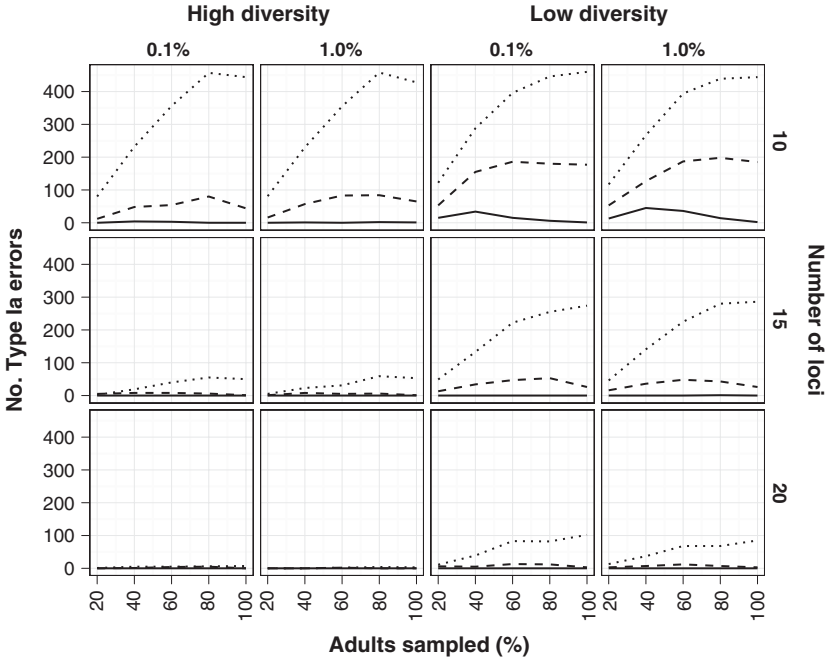


Fig. 2 Susceptibility of three popular methods of parentage analysis to type Ia errors under 60 independent scenarios. Number of false parent–offspring pairs where an offspring was assigned to a parent that was not its true parent when the true parent was in the sample varied with the number of loci (*y*-axis), allelic diversity in two simulated populations (N1000 and N500) level of genotyping error (0.1% and 1.0%). Continuous lines correspond to the results from the full-likelihood method implemented in COLONY v2.0 (Wang 2004), dashed lines are the results from the pairwise-likelihood implemented in FAMOZ (Gerber *et al.* 2003) and dotted lines from the Christie method (Christie 2010a).

methods (Table S5, Supporting information). This included a small but significant interaction with allelic diversity (2.6%: $F_{1,54} = 30.6$, $P < 0.001$) for the Christie method and a small but significant interaction with the number of loci (1.2%: $F_{1,55} = 9.8$, $P < 0.01$) for the pairwise-likelihood method. In contrast, proportion of sampled parents explained ~17.6% of the overall variance in accuracy of the full-likelihood method and showed no significant interaction with other variables.

Overall, the presence of genotyping error had negligible impact on the accuracy of assignments (Fig. 1 and Tables S3–S5, Supporting information). For each method, we compared the average accuracy in the high- and low-diversity data sets with either 0.1% or 1% genotyping error. Overall, a 10-fold increase in genotyping error resulted in a 2–3% reduction in accuracy for the Christie methods and <1% reduction for the pairwise- and full-likelihood methods.

Trends in error types

The relative accuracy of each method was reflected in their susceptibility to type Ia, type Ib and type II errors (Box 1), and although incomplete sampling of candidate parents was not highly descriptive of the variance in overall accuracy (<10% for all methods), it was highly significant ($P < 0.001$) and defined clear trends in error rates, irrespective of the number of loci and allelic diversity.

For both the Christie method and pairwise-likelihood methods, the number of type Ia errors (falsely assigning to a parent when the true parent is in the sample) increased as the proportion of candidate parents in the sample increased (Fig. 2). In most scenarios investigated, the Christie method was the most susceptible to type Ia errors, which represented 45% of all false assignments. The susceptibility of the pairwise-likelihood approach to type Ia errors, though not as sensitive as the Christie method with fewer errors representing 38% of all errors overall, also increased with increasing proportion of the adult sample. Whilst the number of type Ia errors appears to asymptote when the proportion of adults reached 60% and 80% for the pairwise-likelihood approach and Christie method, respectively, it did not necessarily decrease beyond that point.

Furthermore, both the Christie methods and the pairwise-likelihood method were also susceptible to type Ib errors (falsely assigning to a parent when the true parent is not in the sample), with the number of errors decreasing as the proportion of sampled parents increased (Fig. 3). These were the most common forms of error for the pairwise-likelihood method (39%) and, whilst the overall trend and susceptibility were similar between the two approaches, these were the least likely error for the Christie method (15% of errors overall). Given the accuracy of the full-likelihood method, clear trends were not easily identified. Low number and diversity of loci did appear to increase the susceptibility of this approach to both type Ia and type Ib errors, representing 20% and 57% of all errors, respectively. Both error types decreased with over 40% of the adult population sampled.

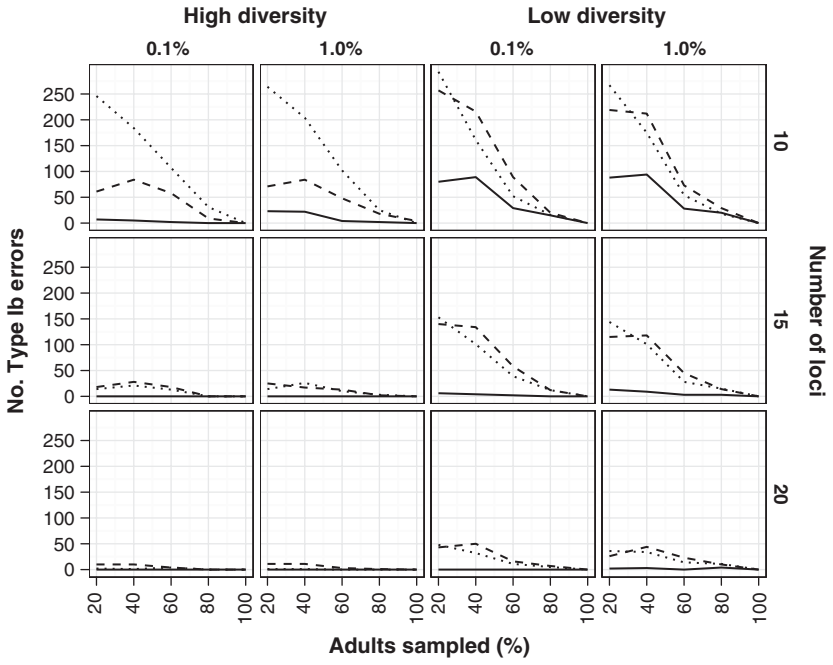


Fig. 3 Susceptibility of three popular methods of parentage analysis to type Ib errors under 60 independent scenarios. Number of false parent–offspring pairs where an offspring was assigned to a parent that was not its true parent when the true parent was not in the sample varied with the number of loci (*y*-axis), allelic diversity in two simulated populations (N1000 and N500) level of genotyping error (0.1% and 1.0%). Continuous lines correspond to the results from the full-likelihood method implemented in COLONY v2.0 (Wang 2004), dashed lines are the results from the pairwise-likelihood implemented in FAMOZ (Gerber *et al.* 2003) and dotted lines from the Christie method (Christie 2010a).

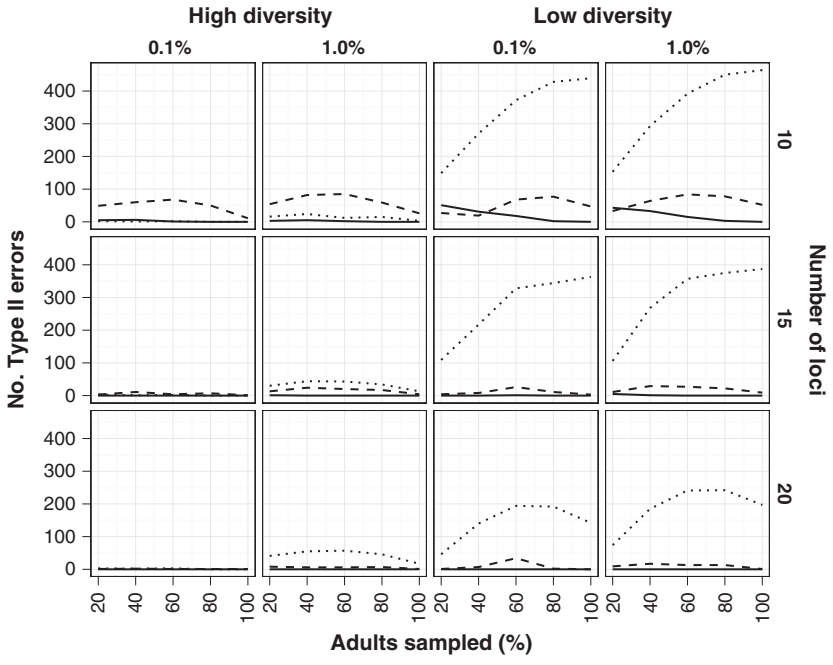


Fig. 4 Susceptibility of three popular methods of parentage analysis to type II errors under 60 independent scenarios. Number of false parent–offspring pairs where an offspring was not assigned when the true parent was in the sample varied with the number of loci (*y*-axis), allelic diversity in two simulated populations (high and low diversity) level of genotyping error (0.1% and 1.0%). Continuous lines correspond to the results from the full-likelihood method implemented in COLONY v2.0 (Wang 2004), dashed lines are the results from the pairwise-likelihood implemented in FAMOZ (Gerber *et al.* 2003) and dotted lines from the Christie method (Christie 2010a).

The occurrence of type II errors (falsely excluding a parent when it was in the sample) remained low for both the pairwise-likelihood and full-likelihood methods, representing 22% and 23% of false assignments overall (Fig. 4). However, these increased sharply with sample size for the Christie method under scenarios with low allelic diversity, representing 40% of false assignments overall. Furthermore, high genotyping error resulted in an increase in type II errors for this method.

Discussion

This study evaluates the performance of three popular approaches to parentage analysis using microsatellite loci in open populations. In these simulated scenarios, we were able to capture a wide diversity of conditions that are commonly encountered in parentage studies and identified key factors for the identification of true parent–offspring pairs in natural populations. We also

identify the three main error types that lead to false assignments. Our results show that with many highly diverse loci, all three methods investigated identified true parent–offspring pairs with high levels of accuracy. However, as we reduced the number and allelic diversity of loci and the proportion of parents sampled, the performance of each method responded differently. In general, accuracy declined with reduced number of loci and allelic diversity, whilst the response to the proportion of population sampled and effects of genotyping error varied with each method. In these simulated settings, the full-likelihood approach implemented in COLONY (Wang 2004; Jones & Wang 2010a), consistently outperformed both the pairwise-likelihood method implemented in FAMOZ (Gerber *et al.* 2003) and the Christie method (Christie 2010a), which was subject to the most erroneous assignments.

Accounting simultaneously for parent–offspring pairs and full- and half-sibs clearly increases the accuracy of assignments for the full-likelihood approach implemented in COLONY (Wang 2007; Walling *et al.* 2010). Whilst the inclusion of full-sibs in our simulated data sets was necessary to reduce the computational demands of this method, these may not be present at such frequencies in natural populations. Furthermore, low allelic diversity or the absence of many candidate parents makes the identification of family clusters much more difficult. Whilst the most informative data sets (20 loci with 100% sampled parents and high allele diversity) took several hours to complete, the least informative data sets (10 loci with 20% sampled parents and low allele diversity) took up to 4 months to complete (single run with high precision on a single CPU). For larger natural populations with mixed generations and complex genealogical relationships, it would take considerably longer. Consequently, the performance of COLONY may be reduced if the presence of large family groups is infrequent, a common characteristic of both terrestrial and marine systems (Selkoe *et al.* 2006; Buston *et al.* 2009). How the presence of full-sibs in the sample, as candidate parents or as offspring, affects performance remains unclear and requires further investigation. The computation time remains the major drawback of this method, and thus, its application may be restricted to studies with small sample sizes. However, during the development of this study, a new likelihood method was released (Wang 2012) that is less computationally demanding than the full-likelihood method and may overcome this limitation.

Although the pairwise-likelihood method implemented in FAMOZ (Gerber *et al.* 2003) was sensitive to the number and allelic diversity of loci used in the analysis, it is a good compromise to the full-likelihood approach. Whilst it did not perform as well, it is well suited for parentage studies in large natural populations where knowledge of biological or demographic parameters is limited or unavailable, where sample sizes are large or where the number and diversity of loci are limited. Furthermore, prior knowledge of the proportion of candidate parents sampled did not appear to be an important factor in determining true parent–offspring pairs. The pairwise-likelihood method is also far less computationally intensive, with each run taking only minutes to complete on a standard laptop computer. The flexibility of FAMOZ allows for a broad range of applications and has made it an attractive approach to investigate mating patterns and dispersal in a variety of taxa where demographic parameters are often difficult to obtain.

The performance of the Christie method (Christie 2010a) was clearly affected by the number and allelic diversity of molecular markers chosen in these simulated scenarios. However, provided that enough highly diverse markers are available, the accuracy of this method increases substantially. Low allelic diversity combined with large sample sizes increases the probability of false parent–offspring pairs in the sample and would explain the susceptibility of this method to type II errors in low-diversity data sets (Christie 2010a). Setting a threshold whereby putative assignments are only accepted if the probability of false assignments is <0.10 or 0.05 was attempted to reduce the number of false positives; however, the increase in false negatives outweighed the benefits and did not increase the overall accuracy of assignments. Overall, we found the approach computationally intensive, especially in scenarios where the number and diversity of loci were low, perhaps due to the standardized number of simulations we chose. One potential constraint of this approach is the inability to identify parent pairs, limiting its application for ecological studies if no demographic or mating information is available. Nevertheless, this method is well suited for situations where only small proportions of large populations can be sampled (e.g. Christie *et al.* 2010b) and has been successfully applied to infer reproductive success in a captive-breeding programme (Christie *et al.* 2011, 2012).

In natural populations where exhaustive sampling is prohibitive, variation in the proportion of sampled parents can have a significant impact on the accuracy of parentage reconstructions (Nielsen *et al.* 2001; Hadfield *et al.* 2006; Koch *et al.* 2008). Our results show that sampling higher proportions of the population decreases the likelihood of falsely assigning to a parent when the true parent was not in the sample (type Ib). This is simply because more true parents are present in the sample. On the other hand, sampling higher proportions of the population increases the likelihood falsely assigning

to a parent (type Ia) or falsely excluding a parent (type II) when the true parent was in fact in the sample. Sampling larger proportions of adults leads to exponential increases in the number of possible pairwise comparisons and limited genetic information leads to erroneous assignments. Nevertheless, our results showed that for all methods, increasing the number and allelic diversity of loci reduced the effects of incomplete sampling to the point where they became negligible.

Depending on the objectives of the study, different types of errors will have different consequences on the interpretation of the results. For example, if the objective is simply to assign offspring to a population or a group of individuals (e.g. to estimate self-recruitment rates at the population level), type Ia errors will have little bearing on the conclusions of a study because they will not affect the proportion of assignments. On the other hand, if one was to measure individual reproductive success (e.g. Rodriguez-Munoz *et al.* 2010; Beldade *et al.* 2012), any error type can have adverse consequences and assignments may not necessarily reflect true ecological processes. Regardless of the method used, performing simulations to estimate different error rates could help to identify the number of markers required to address specific questions. Striking a balance will be necessary to achieve the best performance or satisfy the objectives of a given parentage study.

Conclusions

This study highlights how the number and diversity of loci, the proportion of candidate parents sampled and the level of genotyping error can affect the accuracy of parentage assignments in three common methods of parentage analysis. Within the range of values simulated, COLONY outperformed FAMOZ, which outperformed the Christie method. However, with 20 or more highly polymorphic loci, all methods could be applied with confidence, though which method is most suitable is likely to depend on the size of the data set and the size of the population investigated. When using fewer loci or less diverse loci, it is vital to be aware of the potential for assignments errors and the nature of these errors when choosing which method to apply. Parentage studies in natural populations are a challenging endeavour, and obtaining accurate assignments is crucial to obtaining accurate representations of ecological processes. Whilst most studies will seek to minimize false assignments, a compromise between the cost of developing and processing a large number of loci and sampling effort is often necessary. Obtaining larger sample sizes of potential adults obviously increases the number of possible assignments. However, we found that increasing the number of loci or selecting loci with greater allelic diversity can compensate for incomplete sampling of the parental population and still achieve high levels of accuracy.

Acknowledgements

We would like to thank Jinliang Wang for kindly sharing the Fortran source code of COLONY2, Samar Aseeri and Dodi Heryadi from the KAUST Supercomputing Laboratory for compiling the program to enable parallel computation on Mac OSX and Wayne Mallett for assistance with the high-performance computing facility at James Cook University. We are grateful for comments from Glenn Almany and would also like to thank Peter Buston and two anonymous reviewers for insightful reviews. HBH was supported by a James Cook University Post-Graduate Research Scholarship in cotutelle with the Ecole Pratique des Hautes Etude, Université de Perpignan. PSA was supported by a Post-Doctoral Research Fellowship from King-Abdullah University of Science and Technology.

References

Almudevar A, Anderson EC (2011) A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. *Molecular Ecology Resources*, **12**, 164–178.

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.

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

Balloux F (2001) EASYPOP (Version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.

Beldade R, Holbrook SJ, Schmitt RJ, Planes S, Malone D, Bernardi G (2012) Larger female fish contribute disproportionately more to self-replenishment. *Proceedings of the Royal Society. B, Biological Sciences*, **279**, 2116–2121.

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

Berumen ML, Almany GR, Planes S, Jones GP, Saenz-Agudelo P, Thorrold SR (2012) Persistence of self-recruitment and patterns of larval connectivity in a marine protected area network. *Ecology and Evolution*, **2**, 444–452.

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

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays*, **28**, 1040–1050.

Buston PM, Bogdanowicz SM, Wong A, Harrison RG (2007) Are clownfish groups composed of close relatives? An analysis of microsatellite DNA variation in *Amphiprion percula*. *Molecular Ecology*, **16**, 3671–3678.

Buston PM, Fauvelot C, Wong MY, Planes S (2009) Genetic relatedness in groups of the humbug damselfish *Dascyllus*

aruanus: small, similar-sized individuals may be close kin. *Molecular Ecology*, **18**, 4707–4715.

Christie MR (2010a) Parentage in natural populations: novel methods to detect parent-offspring pairs in large data sets. *Molecular Ecology Resources*, **10**, 115–128.

Christie MR, Tissot BN, Albins MA *et al.* (2010b) Larval connectivity in an effective network of marine protected areas. *PLoS ONE*, **5**, e15715.

Christie MR, Marine ML, Blouin MS (2011) Who are the missing parents? Grandparentage analysis identifies multiple sources of gene flow into a wild population. *Molecular Ecology*, **20**, 1263–1276.

Christie MR, Marine ML, French RA, Blouin MS (2012) Genetic adaptation to captivity can occur in a single generation. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 238–242.

Crawley MJ (2007) *The R Book*. Wiley & Sons Ltd, Chichester, UK.

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.

Ford MJ, Hanson MB, Hempelmann JA *et al.* (2011) Inferred paternity and male reproductive success in a killer whale (*Orcinus orca*) population. *Journal of Heredity*, **102**, 537–553.

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

Garcia C, Arroyo JM, Godoy JA, Jordano P (2005) Mating patterns, pollen dispersal, and the ecological maternal neighbourhood in a *Prunus mahaleb* L. population. *Molecular Ecology*, **14**, 1821–1830.

Garcia C, Jordano P, Godoy JA (2007) Contemporary pollen and seed dispersal in a *Prunus mahaleb* population: patterns in distance and direction. *Molecular Ecology*, **16**, 1947–1955.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines: recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources*, **11**, 1093–1101.

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.

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

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591–611.

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.

Harrison XA, Tregenza T, Inger R *et al.* (2010) Cultural inheritance drives site fidelity and migratory connectivity in a long-distance migrant. *Molecular Ecology*, **19**, 5484–5496.

Harrison HB, Williamson DH, Evans RD *et al.* (2012a) Larval export from marine reserves and the recruitment benefit for fish and fisheries. *Current Biology*, **22**, 1023–1028.

Harrison HB, Saenz-Agudelo S, Planes S, Jones GP, Berumen ML (2012b) Data from: relative accuracy of three common methods of parentage analysis in natural populations. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.2ht96

Herbinger CM, O'Reilly PT, Verspoor E (2006) Unravelling first-generation pedigrees in wild endangered salmon populations using molecular genetic markers. *Molecular Ecology*, **15**, 2261–2275.

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

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

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

Jones OR, Wang J (2010a) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551–555.

Jones OR, Wang J (2010b) Molecular marker-based pedigrees for animal conservation biologists. *Animal Conservation*, **13**, 26–34.

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

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, **10**, 6–30.

Jordano P, Garcia C, Godoy JA, García-Castaño JL (2007) Differential contribution of frugivores to complex seed dispersal patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 3278–3282.

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.

Kanno Y, Vokoun JC, Letcher BH (2011) Sibship reconstruction for inferring mating systems, dispersal and effective population size in headwater brook trout (*Salvelinus fontinalis*) populations. *Conservation Genetics*, **12**, 619–628.

Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology and Evolution*, **17**, 230–241.

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

Magrath MJL, Vedder O, van der Velde M, Komdeur J (2009) Maternal effects contribute to the superior performance of extra-pair offspring. *Current Biology*, **19**, 792–797.

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

Morrissey MB, Wilson AJ (2005) The potential costs of accounting for genotyping errors in molecular parentage analyses. *Molecular Ecology*, **14**, 4111–4121.

Nathan R, Perry G, Cronin JT, Strand AE, Cain ML (2003) Methods for estimating long-distance dispersal. *Oikos*, **103**, 261–273.

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

Nielsen JF, English S, Goodall-Copestake WP *et al.* (2012) Inbreeding and inbreeding depression of early life traits in a cooperative mammal. *Molecular Ecology*, **21**, 2788–2804.

Nussey DH, Coltman DW, Coulson T *et al.* (2005) Rapidly declining fine-scale spatial genetic structure in female red deer. *Molecular Ecology*, **14**, 3395–3405.

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.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society. B, Biological Sciences*, **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.

Piyapong C, Butlin RK, Faria JJ, Scruton KJ, Wang J, Krause J (2010) Kin assortment in juvenile shoals in wild guppy populations. *Heredity*, **106**, 749–756.

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 of the United States of America*, **106**, 5693–5697.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reeve HK, Westneat DF, Noon WA, Sherman PW, Aquadro CF (1990) DNA "fingerprinting" reveals high levels of inbreeding in colonies of the eusocial naked mole-rat. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 2496–2500.

Richardson DS, Jury FL, Blaakmeer K, Komdeur J, Burke T (2001) Parentage assignment and extra-group paternity in a cooperative breeder: the Seychelles warbler (*Acrocephalus sechellensis*). *Molecular Ecology*, **10**, 2263–2273.

Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology*, **9**, 1195–1204.

Rodriguez-Munoz R, Bretman A, Slate J, Walling CA, Tregenza T (2010) Natural and sexual selection in a wild insect population. *Science*, **328**, 1269–1272.

Saenz-Agudelo P, Jones GP, Thorrold SR, Planes S (2009) Estimating connectivity in marine populations: an empirical evaluation of assignment tests and parentage analysis at different spatial scales. *Molecular Ecology*, **18**, 1765–1776.

Saenz-Agudelo P, Jones GP, Thorrold SR, Planes S (2011) Connectivity dominates larval replenishment in a coastal reef fish metapopulation. *Proceedings of the Royal Society. B, Biological Sciences*, **278**, 2954–2961.

Saenz-Agudelo P, Jones GP, Thorrold SR, Planes S (2012) Patterns and persistence of larval retention and connectivity in a marine fish metapopulation. *Molecular Ecology*, **21**, 4695–4705.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.

Selkoe KA, Gaines SD, Caselle JE, Warner RR (2006) Current shifts and kin aggregation explain genetic patchiness in fish recruits. *Ecology*, **87**, 3082–3094.

Slavov GT, Leonardi S, Adams WT, Strauss SH, DiFazio SP (2010) Population substructure in continuous and fragmented stands of *Populus trichocarpa*. *Heredity*, **105**, 348–357.

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

Walling CA, Pemberton JM, Hadfield JD, Kruuk LEB (2010) Comparing parentage inference software: reanalysis of a red deer pedigree. *Molecular Ecology*, **19**, 1914–1928.

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

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

Wang J (2010) Effects of genotyping errors on parentage exclusion analysis. *Molecular Ecology*, **19**, 5061–5078.

Wang J (2012) Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics*, **191**, 183–194.

Wang J, Santure AW (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, **181**, 1579–1594.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Wilson AJ, Ferguson MM (2002) Molecular pedigree analysis in natural populations of fishes: approaches, applications, and practical considerations. *Canadian Journal of Fisheries and Aquatic Sciences*, **59**, 1696–1707.

H.B.H., P.S.-A., S.P., G.P.J. and M.L.B. designed study, H.B.H. and P.S.-A. performed the parentage analyses, H.B.H. and P.S.-A. wrote the original manuscript and all authors contributed to revisions. All authors have a general interest in coral reef ecology and measuring larval dispersal and connectivity in coral reef fishes.

Data accessibility

Simulated data sets and R scripts deposited in the Dryad Repository: http://dx.doi.org/10.5061/dryad.2ht96.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Characteristics of 20 simulated loci for the N1000 high diversity population.

Table S2 Characteristics of 20 simulated loci for N500 low diversity population.

Table S3 Accuracy of three popular methods for parentage analysis in open populations and assignment errors for the N1000 simulated dataset.

Table S4 Accuracy of three popular methods for parentage analysis in open populations and assignment errors for the N500 simulated dataset.

Table S5 Results from the Generalized Linear Model on the accuracy of three methods of parentage analysis.