MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 5731–5737

NEWS AND VIEWS

COMMENT

Bayesian parentage analysis reliably controls the number of false assignments in natural populations

MARK R. CHRISTIE

Department of Zoology, Oregon State University, Corvallis, OR 97331-2914, USA

Parentage analysis in natural populations is a powerful tool for addressing a wide range of ecological and evolutionary questions. However, identifying parent–offspring pairs in samples collected from natural populations is often more challenging than simply resolving the Mendelian pattern of shared alleles. For example, large numbers of pairwise comparisons and limited numbers of genetic markers can contribute to incorrect assignments, whereby unrelated individuals are falsely identified as parent–offspring pairs. Determining which parentage methods are the least susceptible to making false assignments is an important challenge facing molecular ecologists. In a recent paper, Harrison *et al.* (2013a) address this challenge by comparing three commonly used parentage methods, including a Bayesian approach, in order to explore the effects of varied proportions of sampled parents on the accuracy of parentage assignments. Unfortunately, Harrison *et al.* made a simple error in using the Bayesian approach, which led them to incorrectly conclude that this method could not control the rate of false assignment. Here, I briefly outline the basic principles behind the Bayesian approach, identify the error made by Harrison *et al.*, and provide detailed guidelines as to how the method should be correctly applied. Furthermore, using the exact data from Harrison *et al.*, I show that the Bayesian approach actually provides greater control over the number of false assignments than either of the other tested methods. Lastly, I conclude with a brief introduction to SOLOMON, a recently updated version of the Bayesian approach that can account for genotyping error, missing data and false matching.

Keywords: exclusion, kinship, parentage, pedigree, relatedness, reproductive success

Received 25 January 2013; revised 19 March 2013; accepted 25 March 2013

Determining the parentage of individuals collected from wild populations is important for addressing a broad range of ecological and evolutionary questions (Jones & Ardren 2003; Pemberton 2008). A primary challenge confronting successful parentage analysis in natural populations is to control the number of false assignments (Jamieson & Taylor 1997; Christie 2010), which can occur when individuals that are not part of a parent–offspring relationship are incorrectly assigned as such. False assignments can readily occur in data sets collected from natural populations due to the large number of pairwise comparisons that must be made between putative parents and offspring. For example, if 100 adults and 100 juveniles are collected from the wild without any supplementary demographic or observational data, then 10 000 pairwise comparisons are required (i.e. each adult must be compared with each juvenile). With limited numbers of genetic loci, these large numbers of comparisons can result in pairs of individuals sharing alleles across all loci by chance alone, falsely mimicking the patterns of alleles shared between parents and their offspring. False assignments are a serious concern for most parentage studies because they can lead to incorrect conclusions regarding reproductive success, individual fitness and pedigree reconstruction (Marshall *et al.* 1998; Araki & Blouin 2005; Kalinowski *et al.* 2007). Thus, a primary objective of many parentage methods is to minimize the number of false assignments (see Jones & Ardren 2003 and Jones *et al.* 2010 for a review of parentage methods). In an attempt to identify which methods can identify the most parent–offspring pairs while minimizing the number of false assignments, Harrison *et al.* (2013a) applied three different parentage methods to simulated data sets. The authors employed a kinship program, COLONY (Jones & Wang 2010), a likelihood-based approach, FAMOZ (Gerber *et al.* 2003), and a Bayesian parentage method (Christie 2010) that was suggested to have a high rate of false assignment. Here, I show that Harrison *et al.* used incorrect settings for the Bayesian approach and demonstrate that if more appropriate settings are used, then this approach actually performs better than the others at controlling the number of false assignments.

The Bayesian parentage method (Christie 2010) was initially developed to identify parent–offspring pairs in data sets where a low proportion of parents and offspring were sampled. For example, the majority of marine larvae are miniscule and are nearly impossible to directly track in their ocean environment. As such, parentage analyses can be a useful way to determine how far and to what extent larvae are dispersing, a result that has important implications for the successful design and implementation of marine-protected areas (Planes *et al.* 2009; Christie *et al.* 2010a,b; Harrison *et al.* 2012). When sampling from large natural populations, the probability of collecting any parent–offspring pairs can be low given the small size of the sample relative to the size of the population. Thus, in

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

some cases, there may not be any true parents or offspring within the sample, such that it is vital to accurately control the rate of false assignments.

With that end in mind, the approach presented in Christie (2010) first calculates a prior probability of any putative pair being false within a data set. Within this framework, a putative pair is any pair of individuals that share at least one allele at all loci (see Christie *et al.* 2013 for a more flexible approach). The prior is calculated directly from the genetic data and is obtained by dividing the expected number of pairs that share alleles at all loci by chance alone by the observed number of putative pairs. For example, if the number of pairs expected to share an allele at all loci by chance was equal to 300 and the observed number of putative pairs in the data set equalled 1000, then the prior would be equal to 0.30. Thus, the probability that a randomly selected putative pair in the data set is false would equal 30%. This prior is next incorporated into Bayes' theorem to calculate the posterior probability for each individual putative parent–offspring pair. Even if the prior probability of a randomly selected pair being false is high, the posterior probability for an individual pair may be low because true parent–offspring pairs often share alleles that are less common than those observed in pairs that share alleles by chance (See Box 1 for a brief summary of the method).

Box 1 Bayesian parentage analysis

To identify parent–offspring pairs, Bayes' theorem is employed to determine the posterior probability of a putative parent–offspring pair being false given the frequencies of shared alleles. Briefly, the method takes into account allele frequencies such that pairs that share common alleles are considered much less likely to be true than are pairs that share rare alleles. In accordance with Mendelian expectation, each parent–offspring pair will share at least one allele across all loci. If a limited number of loci are employed, then pairs of individuals can share alleles by chance alone. The rate of false matching for a given marker set increases exponentially with a linear increase in sample size (Christie 2010). The first step is to calculate a prior equal to the probability of any randomly selected putative pair sharing alleles by chance:

$$\Pr(\phi) = \frac{F\text{pairs}}{N\text{putative}} \qquad (eqn\ 1)$$

where *F*pairs equals the expected number of false parent–offspring pairs and *N*putative equals the total number of putative parent–offspring pairs. Thus, if a data set was expected to contain 10 pairs that shared alleles by chance, but was observed to contain 100 pairs, then Pr(ϕ) would equal 0.1. Calculating these values requires simulations based on the genotype data (See Methods in Christie *et al.* 2013 for details). In rare cases where the expected number of false pairs is greater than the number of putative pairs, the prior is rounded to 1. Most, if not all, observed false pairs will share common alleles, because the probability of sharing an allele by chance is approximately proportional to the square of the allele frequency. In contrast, the probability that a true parent–offspring pair will share a particular allele is simply proportional to the allele frequency. Therefore, pairs sharing rare alleles are much more likely to be true parent–offspring pairs. Bayes' theorem is invoked to exploit this principle by calculating the probability of a putative parent–offspring pair being false given the frequencies of shared alleles:

$$\Pr(\phi|\lambda) = \frac{\Pr(\lambda|\phi) \cdot \Pr(\phi)}{\Pr(\lambda|\phi).\Pr(\phi) + \Pr(\lambda|\phi^c) \cdot \Pr(\phi^c)} \qquad (eqn\ 2)$$

where Pr(ϕ) is the prior, calculated as described above, and Pr(ϕ^c) is the complement. Pr(λ | ϕ) equals the probability of sharing the observed alleles given that the putative pair in question is false. This value is calculated for each putative pair using simulated multilocus genotypes. Notice that when a putative pair shares the most common alleles across all loci that Pr(λ | ϕ) = 1, and consequently Pr(ϕ | λ) = Pr(ϕ). To calculate Pr(λ | ϕ^c), which is the probability of sharing alleles given that a putative pair is true, the same approach is employed, but the observed allele frequencies are used rather than the frequencies at which alleles were shared. Notice also that when the prior Pr(ϕ) equals 1, the posterior, Pr(ϕ | λ), also equals 1.

For each putative pair in the data set, the Bayesian approach calculates a posterior probability of a pair being false given the frequencies of shared alleles. For example, if a particular putative pair has a posterior probability equal to 0.05, then it has a 5% probability of being false. If there are 100 other putative pairs from the same data set with posterior probabilities equal to 0.05, then five of those pairs will be false on average. The appropriate way to employ the Bayesian parentage method is to define an a priori cut-off value (hereafter: alpha) for the calculated posterior probabilities. It is generally not advisable to accept pairs with a posterior value >0.1. These pairs may not necessarily be false, but as the posterior probability increases, there is an increasing probability that the putative pair is false. Using the same example as above, if alpha is set to 0.1, then 10 of 100 assignments may be false, if alpha is set to 0.5, then 50 of 100 assignments may be false, and if alpha is set to 1, as was done by Harrison *et al.*, then all assignments may be false. For most studies, setting alpha between 0.01 and 0.05 will minimize most false assignments (see Christie *et al.* 2013). In rare cases, slightly lower (e.g. 0.001) or higher (e.g. 0.1) cut-off values may be warranted.

In the paper by Harrison *et al.*, the authors used a two-step process when applying the Bayesian method. First, the authors set $\alpha = 1$, thereby accepting all putative pairs, regardless of the posterior probability. Because the Bayesian approach from Christie (2010) reports every putative pair that matches at all loci, this procedure is identical to performing simple exclusion without allowing for a locus to mismatch (Fig. 1A, orange line). Next, for each offspring, the authors accepted only the assignments with the lowest posterior value (Fig. 1A, blue line). Thus, if an offspring matched to one candidate father with a posterior probability equal to 0.01 and to a different candidate father with a posterior probability of 0.09, only the first assignment was included. However, if a different offspring matched one candidate father with a posterior probability of 0.9 and another candidate father with a posterior probability of 0.95, then the first pair would be accepted, even though the probability of that pair being false equalled 0.9. If an offspring only matched a single candidate parent, it was also accepted, regardless of the posterior probability. Not surprisingly, the number of type I errors (false assignments) reported in Harrison *et al.* for the Bayesian method is quite large because alpha (the cut-off value) was set to 1. Here, I reanalyse the exact data presented in Harrison *et al.* using a more appropriate cut-off value.

Methods

Harrison *et al.* created simulated data sets corresponding to low and high genetic diversity. The low-diversity data sets averaged 10.7 alleles per locus and the high-diversity data sets averaged 14.9 alleles per locus. In the low-diversity data set, 250 simulated males and 250 simulated females were monogamously paired to create four offspring each, resulting in 1000 offspring. In every test data set, all 1000 offspring were included, but the proportion of parents sampled was varied from 20% to 100%. Genotyping errors were introduced at 0.1% and 1.0% in the offspring only. The data used here are identical to those used in Harrison *et al.* (2013a) and are available at dryad (Harrison *et al.* 2013b). I first used the low-diversity data sets as they were the most affected by false assignments (see Fig. 2 of Harrison *et al.* 2013a). Using both the traditional Bayesian approach presented in Christie (2010) and the recently updated approach (Christie *et al.* 2013), I recalculated the number of false assignments after setting a cut-off value equal to 0.01 (a conservative cut-off value because only one in 100 assignments should be false at that level). I also used the exact R script to calculate type I and II errors as in Harrison

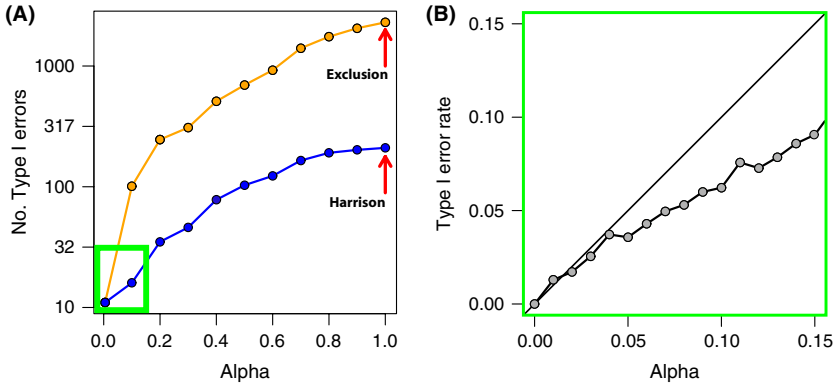


Fig. 1 Effect of varying the maximum allowable posterior probability for correct parentage assignment (alpha) on the number of false assignments (type I error). This data set contains 15 loci and is a low-diversity data set directly from Harrison *et al.* (2013a). The orange line (top line) in panel A shows the number of false assignments for the Bayesian approach as function of alpha (the cut-off value). If alpha is set to 1, then this method is equivalent to using Mendelian incompatibility (Exclusion, red arrow). An alternate approach is to parse the offspring assigned to multiple parents of the same sex and only accept the pair with the lowest posterior probability (blue points). Harrison *et al.*, used this latter approach, but also set $\alpha = 1$ (Harrison, red arrow). The green box in panel A represents the maximum recommended range of posterior values to include in order to minimize false assignments. Panel B illustrates that regardless of the chosen cut-off value, the proportion of type one errors is equal to or less than the chosen cut-off value (alpha). Notice that the type I errors in panel B are expressed as a rate in order to make direct comparisons between alpha and the type I error rate (1:1 line).

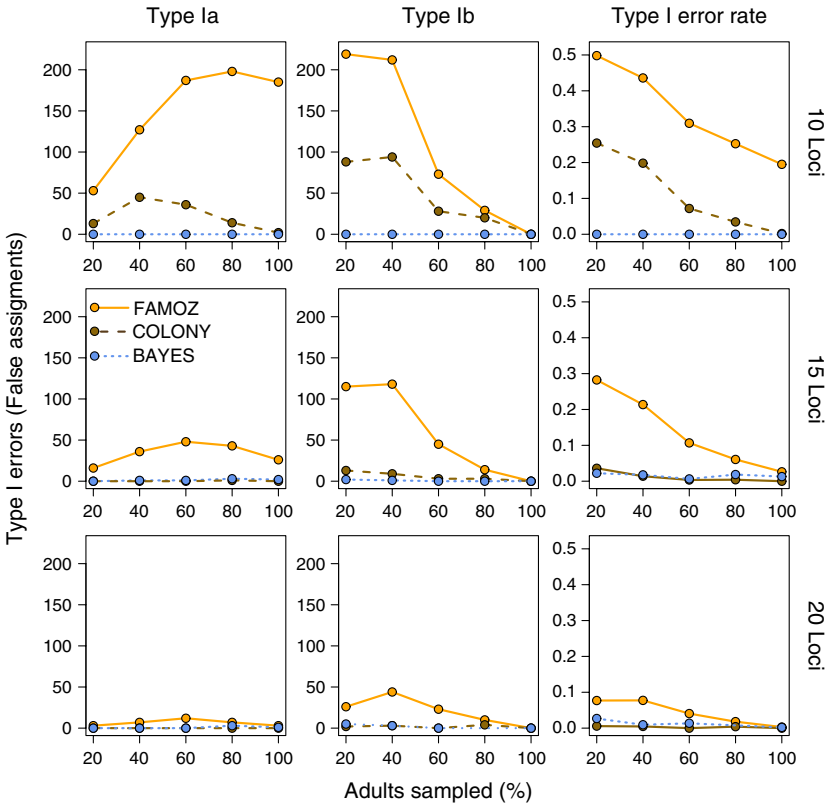


Fig. 2 Using the exact data as in Harrison *et al.*, this figure illustrates the number of type I errors using three parentage methods. The difference between this figure and the data displayed in Harrison *et al.* is that an alpha value of 0.01 was used here (instead of 1). As per Harrison *et al.*, type Ia error is the number of false assignments when the true parent was in the sample. Type Ib error is the number of false assignments when the true parent was not included in the sample. The type I error rate equals the total number of type I errors divided by the total number of assignments. The data here are from the 'low-diversity' data sets. Data for COLONY and FAMOZ are directly from table S4 of Harrison *et al.* (2013a). 'Bayes' refers to the Bayesian parentage approach. Notice that when alpha is set to 0.01, the Bayesian approach has the lowest number of type I errors (and a controllable type I error rate).

et al., which was available as supplementary online material, except that I set $\alpha = 0.01$ instead of 1. As in Harrison *et al.*, type Ia errors were defined as false assignments when the true parents were in the sample, and type Ib be errors were defined as false assignments when the true parents were not included in the sample. I also added both type Ia and type Ib errors together (the total number of false assignments) and divided by the total number of assignments to examine the type I error rate. For all results presented here, I used the same definitions of type I and II errors as presented in Harrison *et al.*, so that the results are directly comparable, but it should be noted that Christie (2010) define these terms differently. Results from FAMOZ and COLONY are directly from tables S3 and S4 of Harrison *et al.* (2013a,b).

As with any statistical approach that controls type I errors, including false discovery rate procedures (Skaug *et al.* 2010), the control of false assignments comes at the expense of the number of correct assignments (Sokal & Rohlf 1994). More technically, the decrease in type I errors (false assignments) is typically accompanied by an increase in type II errors (true parent–offspring pairs that remain unassigned). To illustrate these trade-offs, I used the high-diversity data set with 10 loci and determined the number of correct and false assignments associated with two different cut-off values, 0.01 and 0.05. I chose to use this data set because the Bayesian approach was also suggested by Harrison *et al.* to have a large number of false assignments when 10 loci were used with the high-diversity data sets. For all of the above analyses, I used the data sets with the higher rate of genotyping error (1%) because genotyping errors typically present more challenging conditions for parentage assignment.

Results

Including all posterior values (setting alpha to 1) when using the Bayesian method results in a large number of false assignments (Fig. 1A), as expected. However, when alpha is set to a lower value (e.g. 0.01 or 0.05), the number of false assignments is minimal (Figs 1–3). Furthermore, for the Bayesian approach, the type I error rate typically is less than or equal to the desired cut-off value (Fig. 1B; see also Christie *et al.* 2013). When alpha was set to 0.01, the Bayesian approach had the lowest number of type Ia and type Ib errors (Fig. 2). Furthermore, the overall type I error rate was controlled across all proportions of adults sampled. Comparisons with FAMOZ and COLONY revealed that these programs, at least as used by Harrison *et al.*, result in little control over the type I error rate. For both FAMOZ and COLONY, the type I error rate increased with corresponding decreases in the number of loci used, the overall genetic diversity and the proportion of parents that were sampled. This is a concerning result because naïve users with large data sets and limited numbers of loci may end up with many false assignments. For example, the type I error rate for FAMOZ was 0.5 when 10 loci and only 20% of the adults were sampled (Fig. 2). This translates into half of all parentage assignments being incorrect. One distinct advantage of the Bayesian approach, therefore, is that the user can directly control the number of false assignments.

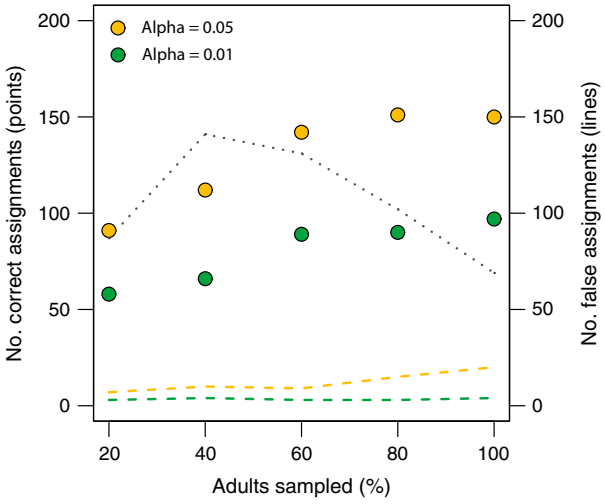


Fig. 3 Using the 'high-diversity' data sets with 10 loci from Harrison *et al.*, this figure illustrates the trade-offs that can arise by varying alpha. When alpha is set at 0.01, there are fewer false assignments (blue dashed line), but also fewer correct assignments (blue points). When alpha is increased to 0.05, there are more false assignments, but also more correct assignments (orange points). Many users may be tempted to set $\alpha = 1$ in order to maximize the number of correct assignments, but this is inadvisable as it may result in a large number of false assignments. The black dotted line illustrates the number of false assignments for FAMOZ (data from table S3 of Harrison *et al.*).

As with any statistical procedure, reducing the number of false assignments comes at the expense of correct assignments. For the high-diversity data sets with 10 loci, the average number of correct assignments was equal to 129.2 when alpha was set to 0.05; however, the average number of correct assignments was reduced to 80 when alpha was set to 0.01 (Fig. 3). This reduction in alpha, however, also corresponded to an average of 9 fewer false assignments, resulting in average of only 3.4 false assignments across data sets. This trade-off is also the explanation as to why Harrison *et al.* set $\alpha = 1$, because as stated in their discussion, the authors were trying to maximize the number of correct assignments (which can maximize their definition of accuracy). In fact, using the definition provided by Harrison *et al.*, accuracy is maximized where the sum of type I and type II errors is lowest. This can result in high numbers of type I errors if there is a multiplicative relationship between type I and type II errors. As such, it is best to report type I and II errors separately. The results presented in figures two and three highlight a feature of this Bayesian parentage method; it tends to be fairly conservative with parentage assignments. As a comparison, FAMOZ averaged 106 false assignments for the high-diversity data sets (Fig. 3), resulting in an average type I error rate of 0.17.

Discussion

For most parentage applications, the ability to control the number of false assignments is a primary concern because all other downstream analyses and conclusions may be biased by incorrect parentage assignments (e.g. estimates of reproductive success). The Bayesian approach performs well at minimizing the number of false assignments when only pairs with small posterior probabilities are accepted. In general, setting a cut-off value between 0.01 and 0.05 represents a good trade-off between maximizing the number of correct assignments while minimizing the number of false assignments (Christie 2010; Christie *et al.* 2013). Depending on the goals of the study, users may want to choose more conservative (i.e. lower, 0.01) or liberal (i.e. higher, 0.05–0.1) cut-off values. A distinct advantage of the Bayesian approach is that it allows one to explicitly control the trade-off. From the data presented in Harrison *et al.* (2013a), the type I error rates for both FAMOZ and COLONY were dependent on the data set being used, but this was not the case with the Bayesian methods.

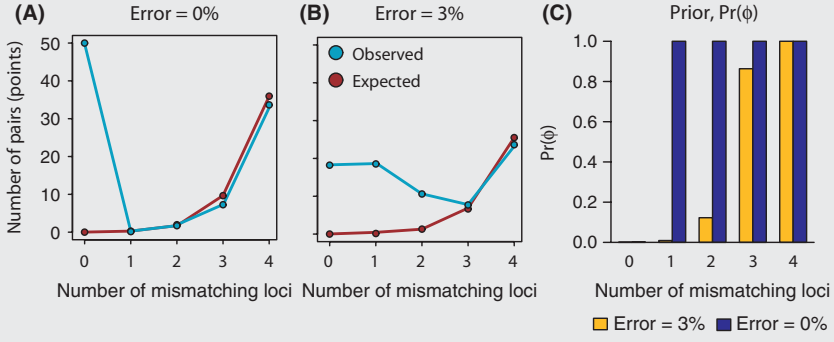
It should be noted that the data sets created by Harrison *et al.* were derived from parents that produced four offspring each. This simulation process resulted in no variation in reproductive success among parents, a result that is unlikely to occur in any natural population. Further comparisons among parentage methods could employ more sophisticated simulations to: (i) explore the effects of variance in reproductive success, (ii) vary the stringency (i.e. criteria for acceptance) between different methods while accounting for the theoretical differences in interpretation, (iii) consider the effects of different mating systems (e.g. monogamy vs. polygamy), (iv) further disentangle the effects of alleles that are shared by descent (with respect to different levels of relationship) vs. those that are simply identical by state and (v) examine the effects of sampling across subdivided populations, which could lead to biased estimates of allele frequencies. Ultimately, there is no one best parentage method for all scenarios and users should carefully decide between methods based on the strengths and weaknesses of their sampling design and their genetic data sets (reviewed in Jones & Ardren 2003).

Recently, the Bayesian parentage methods have been substantially updated and can now be employed via a graphical user interface such that no user knowledge of the R language is required (Christie *et al.* 2013). This improved method now accounts for genotyping error without requiring estimates of the genotyping error rate (See Box 2 for details). The Bayesian procedure has also been expanded to account for scenarios with one known parent or with known parent pairs, which can substantially increase the assignment power (see also Anderson 2012). Furthermore, the required simulation procedure has been revised, such that the run time is several orders of magnitude faster. These sets of R programs have been collated into a single program, now called SOLOMON (Christie *et al.* 2013), and is freely available from CRAN, the R archive network: http://cran.r-project.org. SOLOMON was designed to work with both microsatellites and SNPs. A detailed user manual and installation information can be found at https://sites.google.com/site/parentagemethods/. Because SOLOMON is written in R (R Core Team 2012), the program can be installed on Linux, Mac, and Windows machines.

Box 2 Accounting for genotyping error without estimates of the genotyping error rate

One of the biggest challenges associated with assigning parentage in natural populations is accounting for genotyping errors because they change the expected patterns of Mendelian inheritance. Specifically, errors can cause mismatches such that true parents and offspring no longer share at least one allele across all loci. Polymorphic microsatellite markers are particularly susceptible to genotyping errors due to difficulty scoring, large-allele dropout and stuttering (van Oosterhout *et al.* 2006). Most existing parentage methods can account for genotyping errors, but often require an estimate of the genotyping error rate. Estimating the genotyping error rates can be difficult and often requires regenotyping large subsets of individuals, which can be costly from both a time and monetary perspective (Bonin *et al.* 2004; Pompanon *et al.* 2005). Many studies that employ parentage methods simply assume that their data set has an error rate comparable with other data sets using similar marker types. This approach can be risky because, if the actual error rate is lower or higher than the assumed value, the results may contain fewer true or more false assignments, respectively.

The Bayesian approach presented in Christie (2010) could not explicitly account for genotyping errors. This approach has recently been overhauled to account for genotyping errors without needing any estimates of the genotyping error rate (see Christie *et al.* 2013 for details). This method works by comparing the expected number of false pairs with the observed number of pairs for each number of mismatching loci and uses this ratio as a prior in Bayes' theorem. For example, consider a microsatellite data set with 100 adults, 100 juveniles and 50 true parent–offspring pairs. If the genotyping error rate is 0%, then all 50 pairs will match at all loci (Panel A). In this case, the expected number of false pairs and the observed number of putative pairs will be equal at one or more mismatching loci. This will result in a high prior probability of pairs being false at one or more mismatching loci (Panel C) and subsequently high posterior probabilities for any false pairs with a mismatch. If the genotyping error rate is high at 3%, then a portion of true pairs will mismatch at a handful of loci (Panel B). If the marker set has enough polymorphic loci, then this will reduce the prior at one or more mismatching loci, giving these pairs the possibility of assignment within the Bayesian framework. This approach is effective regardless of whether the mismatch is caused by genotyping error, mutation or missing data. Simulated data in panels A and B are from Christie *et al.* (2013).



The number of type I and II errors will be minimized for all of the above methods by using a large number of polymorphic loci. Determining exactly how many loci are needed before beginning a study can be easily determined with simulations. Such analyses can let researchers know precisely how many loci are needed in order to maximize the number of correct assignments and minimize the number of false assignments. Alternatively, these analyses can let researchers know what sample sizes can be collected for a given marker set and still yield informative results. SOLOMON includes features to create simulated data sets with or without user-specified allele frequencies. These data sets can then be analysed in the program to perform a priori power analyses. The user can define the number of parents, offspring, unrelated individuals, siblings, loci, alleles per locus and genotyping error rate. Hopefully, this feature will be useful for anyone trying to decide how many genetic markers will be necessary for their sample design.

When using SOLOMON, it is important to report the chosen cut-off value (alpha) and to briefly describe why that value was chosen (particularly if alpha is >0.1). In some cases, it may be informative to use both a conservative and liberal cut-off value to analyse the downstream effects. Furthermore, the prior should be reported for all numbers of mismatching loci that include assigned pairs. For example, if some parent–offspring pairs were assigned with 0 mismatching loci and other parent–offspring pairs were assigned with one mismatching locus, then both of the corresponding priors should be reported because they conveniently summarize the exclusionary power of the data set. Lastly, it may be beneficial to report the posterior probability and number of mismatching loci for each parent–offspring pair, particularly if there are

small numbers of assigned pairs. Pairs with lower posterior probabilities are less likely to be false, and it may be useful for readers to examine which pairs were assigned particular probabilities. Lastly, users should be aware that the precision of the posterior probability depends on the number of user-defined simulated data sets and genotypes. High precision is typically achieved by using the default settings, but both the number of simulated data sets and genotypes should be reported to ensure the reproducibility of the posterior probabilities (see user manual for details).

With the continual advancements in next-generation sequencing and reduction in costs associated with genotyping individuals at large numbers of loci, current challenges associated with parentage analysis of natural populations may soon become obsolete. For example, using hundreds of thousands of markers, the program KING can readily distinguish between first and second order relatives (Manichaikul *et al*. 2010). It seems likely that large numbers of loci, coupled with various relatedness and multivariate approaches, will rapidly elucidate all relationships in samples collected from large natural populations. Nevertheless, until individuals can be economically genotyped at many thousands of loci, it remains important to control for false assignments when employing parentage analyses. Here, I have demonstrated that, when used correctly, a Bayesian parentage approach performs well at controlling the number of false assignments. More importantly, the statistical framework allows the user to explicitly control the trade-off between type I and II errors, as needed for the goals of any particular study. The recent advances in incorporating genotyping error and missing data, combined with a user-friendly interface, will hopefully make the Bayesian approach presented in SOLOMON more widely applicable.

Acknowledgements

The author wishes to thank Michael Blouin, Jacob Tennessen and three anonymous reviewers for helpful comments. The author also wishes to thank Hugo Harrison for correspondence and for readily sharing the data sets used in this manuscript. This research was funded by grants from the Bonneville Power Administration to Michael Blouin.

References

Anderson EC (2012) Large-scale parentage inference with SNPS: an efficient algorithm for statistical confidence of parent pair allocations. *Statistical Applications in Genetics and Molecular Biology*, **11**, 1544–6115.

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.

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

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

Christie MR, Johnson DW, Stallings CD, Hixon MA (2010a) Self-recruitment and sweepstakes reproduction amid extensive gene flow in a coral-reef fish. *Molecular Ecology*, **19**, 1042–1057.

Christie MR, Tissot BN, Albins MA *et al*. (2010b) Larval connectivity in an effective network of marine protected areas. *PLoS ONE*, **5**, e15715. doi: 10.1371/journal.pone.0015715.

Christie MR, Tennessen JA, Blouin MS (2013) Bayesian parentage analysis with systematic accountability of genotyping error, missing data, and false matching. *Bioinformatics*, **29**, 725–732.

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.

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

Harrison HB, Saenz-Agudelo P, Planes S, Jones GP, Berumen ML (2013a) Relative accuracy of three common methods of parentage analysis in natural populations. *Molecular Ecology*, **22**, 1158–1170.

Harrison HB, Saenz-Agudelo P, Planes S, Berumen ML, Jones GP (2013b) Data from: relative accuracy of three common methods of parentage analysis in natural populations. *Dryad Digital Repository*. doi:10.5061/dryad.2ht96.

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

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

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

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

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.

Manichaikul A, Mychaleckyj JC, Rich SS *et al*. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.

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

van Oosterhout C, Weetman D, Hutchinson WF (2006) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes*, **6**, 255–256.

Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society B-Biological Sciences*, **275**, 613–621.

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 Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/.

Skaug HJ, Berube M, Palsboll PJ (2010) Detecting dyads of related individuals in large collections of DNA-profiles by controlling the false discovery rate. *Molecular Ecology Resources*, **10**, 693–700.

Sokal RR, Rohlf FJ (1994) *Biometrym*, 3rd edn. W.H. Freeman, New York, NY.

M.R.C. designed research, analyzed the data, and wrote the paper.

doi: 10.1111/mec.12528