BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 6 2013, pages 725–732
doi:10.1093/bioinformatics/btt039

Genetics and population analysis

Advance Access publication January 30, 2013

Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching

Mark R. Christie*, Jacob A. Tennessen and Michael S. Blouin

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

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: The goal of any parentage analysis is to identify as many parent–offspring relationships as possible, while minimizing incorrect assignments. Existing methods can achieve these ends, but they require additional information in the form of demographic data, thousands of markers and/or estimates of genotyping error rates. For many non-model systems, it is simply not practical, cost-effective or logistically feasible to obtain this information. Here, we develop a Bayesian parentage method that only requires the sampled genotypes to account for genotyping error, missing data and false matches.

Results: Extensive testing with microsatellite and SNP datasets reveals that our Bayesian parentage method reliably controls for the number of false assignments, irrespective of the genotyping error rate. When the number of loci is limiting, our approach maximizes the number of correct assignments by accounting for the frequencies of shared alleles. Comparisons with exclusion and likelihood-based methods on an empirical salmon dataset revealed that our Bayesian method had the highest ratio of correct to incorrect assignments.

Availability: Our program SOLOMON is available as an R package from the CRAN website. SOLOMON comes with a fully functional graphical user interface, requiring no user knowledge about the R programming environment. In addition to performing Bayesian parentage analysis, SOLOMON includes Mendelian exclusion and *a priori* power analysis modules. Further information and user support can be found at https://sites.google.com/site/parentagemethods/.

Contact: christim@science.oregonstate.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 8, 2012; revised on January 4, 2013; accepted on January 22, 2013

1 INTRODUCTION

Accurate parentage assignment and pedigree reconstruction are required to make correct inferences for a broad array of study questions (Pemberton, 2008). Parentage methods span a vast gamut of theoretical approaches from fractional to categorical allocation and simple exclusion to sophisticated likelihood-based approaches (Jones and Ardren, 2003; Jones *et al.,* 2010). One area of parentage analysis that has been largely overlooked is a general Bayesian method for categorical allocation. This void is unfortunate, as additional sampling or field information can be elegantly incorporated as priors into a Bayesian framework (Hadfield *et al.*, 2006). Furthermore, the information present within the genotypic data itself can be used to calculate a prior analogous to a false discovery rate, which can be useful for the challenges associated with parentage analysis. As an illustrative example, consider a typical kinship dataset consisting of seven microsatellite loci and 750 individuals (Rieseberg *et al.*, 2012). In this dataset, a parent and an offspring would share at least one allele across all loci following Mendelian inheritance. However, the probability of two unrelated individuals sharing alleles by chance at all loci is not trivial considering that hundreds of thousands of pairwise comparisons are required. Thus, a primary challenge of parentage analysis in natural populations is to correctly identify the true parent–offspring pairs within a dataset, while simultaneously excluding any pairs that share alleles by chance.

The challenge of parentage analysis is further exacerbated by missing data and genotyping errors, which can erode the parent–offspring 'signal' of sharing at least one allele at all loci (Slate *et al.*, 2000; Vandeputte *et al.*, 2006). Because errors can create an incorrect record of genotypes, true parent–offspring pairs in an empirical dataset may not share an allele at all loci despite that being the Mendelian expectation. Here, we address the challenges associated with parentage analysis by first calculating the prior probability of a dyad sharing an allele across all numbers of mismatching loci. The calculation of this prior (analogous to a false discovery rate) creates a systematic framework for determining how many loci to let mismatch and does not require any estimates of genotyping error. For each putative pair, we next use Bayes' theorem to calculate the posterior probability of a parent–offspring pair being false given the frequencies of shared alleles. Because the probability of sharing common rather than rare alleles is much greater for unrelated pairs, we can compare the frequencies of observed shared alleles to a distribution of alleles shared by unrelated individuals. By combining this information with Bayes' theorem, we can maximize the identification of true parents and offspring in a dataset, while minimizing the number of false assignments. Here, we overhaul the approach of Christie (2010) to (i) account for genotyping error and missing data, (ii) reduce the computational time by up to three orders of magnitude as measured in minutes and (iii) allow for one known parent or for known parent pairs (i.e. known matings), which can substantially increase assignment power. We extensively test this methodology with data drawn from three empirical studies and use an empirical salmon dataset to make comparisons with commonly implemented exclusion and likelihood-based methods.

*To whom correspondence should be addressed.

2 METHODS

We created test datasets of multilocus genotypes with allele frequencies based on the site frequency spectra from three empirical studies. We chose empirical studies featuring three distinct taxonomic groups with two different marker types, SNPs and microsatellites (Table 1). The test datasets were fully characterized, such that we knew all true parents and offspring. For drawing comparisons between methods, we used complete genotype data from a summer-run steelhead (*Oncorhynchus mykiss*) dataset (see details later in the text).

2.1 Bayesian parentage method

To identify true parent–offspring pairs, we used Bayes' theorem to determine the posterior probability of a putative parent–offspring pair being false given the frequencies of shared alleles. For illustrative purposes, we first consider a scenario with no missing data, genotyping error or known parents, although we expand on each of these later in the text. 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 used, then pairs of individuals can share alleles by chance alone. In fact, the rate of false matching increases exponentially with a linear increase in sample size (Christie, 2010). We first calculate a prior equal to the probability of any given putative pair sharing alleles by chance:

$$\Pr(\phi) = \frac{Fpairs}{Nputative} \quad (1)$$

where *Fpairs* equals the expected number of false parent–offspring pairs, and *Nputative* equals the total number of putative parent–offspring pairs. Here, we define a 'false parent–offspring pair' to be a pair of unrelated individuals that share alleles by chance. A 'putative parent–offspring pair' is any pair of individuals that share alleles across all loci and contain all true and false parent–offspring pairs. Thus, if a dataset was expected to contain 10 pairs that shared alleles by chance, but was observed to contain 100 pairs, then $\Pr(\phi)$ would equal 0.1. Estimates for $\Pr(\phi)$ are constrained to range between 0 and 1. To calculate the expected number of false pairs in a dataset, we deviate from the approach presented in Christie (2010) and use simulations rather than allele frequencies. We chose to use simulations because they (i) facilitate the incorporation of genotyping error into a Bayesian framework and (ii) substantially expedite the calculation of the posterior probability.

Table 1. Empirical datasets used to validate the Bayesian parentage method

Symbol	Species	Marker	NL	NA	Max
	European Beech (*Fagus sylvatica*)	μsat	13	11.08	0.66
	Steelhead Trout (*O.mykiss*)	μsat	8	34.88	0.19
	Labrador Retriever (*Canis lupus familiarus*)	SNP	21 115 (200)	2.00	0.98

NL refers to the total number of loci used in the study, NA equals the average number of alleles per locus and Max equals the frequency of the most common allele in the dataset. The retriever dataset had 21 115 SNPs of which 200 were randomly selected. References are as follows: beech (Lander *et al.*, 2011), steelhead (Araki *et al.*, 2007) and retriever (Akey *et al.*, 2010).

To determine the expected number of false pairs, we first calculate allele frequencies across all loci. For each locus separately, we calculate genotype frequencies in accordance with Hardy–Weinberg equilibrium (HWE) and create a pool of genotypes where the rarest genotype occurs at least 100 times. We next create simulated genotypes by sampling from this pool a number of individuals equal to the number genotyped in the empirical dataset (randomly assigning individuals as adults and juveniles). We then make all pairwise comparisons between adults and juveniles and calculate the number of times each allele is shared. If a shared allele is homozygous in an individual, then that allele is only counted once. If an adult and a juvenile are heterozygous for the same alleles, then only the rarer of the two alleles is counted. The number of times that an allele is not shared between an adult and a juvenile is also recorded. The user may choose how many simulated datasets per locus that they wish to use, although we recommend a minimum of 100 simulations for SNPs and 1000 simulations for microsatellites to maximize precision for the posterior probability (Supplementary Table S1). In the simulations, we examine each locus separately to expedite the calculation and reduce the amount of memory allocated by R (R Core Team, 2012).

We next create a user-defined number of multilocus 'genotypes' by using the output of the simulations ('simulated genotypes'). Assuming independence across loci, we sample alleles at each locus by the average frequencies that they were observed to be shared between two unrelated individuals. Included in the sampling process is a dummy variable that represents the frequency of dyads that did not share an allele. This process simultaneously creates a distribution of frequencies of alleles shared among false parent–offspring pairs, while also creating a distribution of the number of false pairs that share at least one allele at 0, 1, 2 . . . L loci, where L equals the total number of genotyped loci. We calculate the expected number of false pairs as

$$Fpairs = NLsim \cdot n_1 \cdot n_2 \quad (2)$$

where *NLsim* equals the frequency of the simulated multilocus genotypes that shared at least one allele at all loci and n_1 and n_2 equal the empirical sample sizes of the adults and juveniles. After *Fpairs* is calculated, the number of observed putative pairs (*Nputative*) is calculated using Mendelian incompatibility and used to calculate the prior, $\Pr(\phi)$.

Most, if not all, observed false pairs will share common alleles, as 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. We exploit this principle by using Bayes' theorem to calculate the probability of a putative parent–offspring pair being false given the frequencies of shared alleles:

$$\Pr(\phi|\lambda) = \frac{\Pr(\lambda|\phi) \cdot \Pr(\phi)}{\Pr(\lambda|\phi) \cdot \Pr(\phi) + \Pr(\lambda|\phi^C) \cdot \Pr(\phi^C)} \quad (3)$$

where $\Pr(\phi)$ is calculated as described earlier in the text, and $\Pr(\phi^C)$ is the complement. $\Pr(\lambda|\phi)$ equals the probability of sharing the observed alleles given that the putative pair in question is false. We calculate this value for each putative pair using the multilocus 'genotypes' where each locus contains values equal to the frequencies of alleles shared by false pairs. To create a distribution of frequencies of shared alleles among false parent–offspring pairs, we multiply these values across all loci ('false-pair products'). We similarly calculate the product of the shared allele frequencies among all putative parent–offspring pairs ('putative-pair products'). To calculate $\Pr(\lambda|\phi)$ for each putative pair, we count the number of false-pair products that were less than or equal to the observed putative-pair product and divide by the total. Notice that when a putative pair shares the most common alleles across all loci that $\Pr(\lambda|\phi) = 1$, and consequently $\Pr(\phi|\lambda) = \Pr(\phi)$. To calculate $\Pr(\lambda|\phi^C)$, which is the probability of sharing alleles given that a putative pair is true, we used the same approach, but we use the observed allele frequencies rather than the frequencies at which alleles were shared.

2.2 Genotyping error

Using the simulations, we calculate $\Pr(\phi)$ for every number of mismatching loci (0,1,..,L). When $\Pr(\phi)$ equals unity, the expected number of false pairs equals the total number of putative pairs within the dataset. Mathematically speaking, when the prior $\Pr(\phi)$ equals one, the posterior, $\Pr(\phi|\lambda)$, also equals 1. Consequently, when $\Pr(\phi)$ is equal to 1, there is insufficient power to distinguish between true and false parent–offspring pairs (Fig. 1). In high-power datasets, the expected number of false parent–offspring pairs will be low for the first several mismatching loci. SOLOMON calculates $\Pr(\phi)$ for every number of mismatching loci and calculates $\Pr(\phi|\lambda)$ for all putative pairs where $\Pr(\phi)$ is <1. Notice that the number of loci allowed to mismatch depends on the genotyping error rate and the power of the dataset. If a dataset has no genotyping error, then $\Pr(\phi)$ will equal 1 when allowing a single locus to mismatch because the expected number of false pairs will equal the total number of putative pairs (i.e. all true pairs will not mismatch at a locus, and, consequently, all putative pairs will be false pairs for a positive number of mismatching loci). Conversely, if the same dataset has a high rate of genotyping error, then there will be more true pairs mismatching at a single locus. When there are more true pairs, the total number of putative pairs will increase, and $\Pr(\phi)$ will be <1 provided that the expected number of false pairs is low, and the locus will be allowed to mismatch (Fig. 1). Thus, the number of loci allowed to mismatch is dictated by the genotyping error rate and the expected number of false pairs. In the aforementioned framework, missing data are simply treated as a mismatch, as there is no way to know whether a putative pair would have shared an allele where an individual is missing data. Null alleles can be accounted for by loading in adjusted

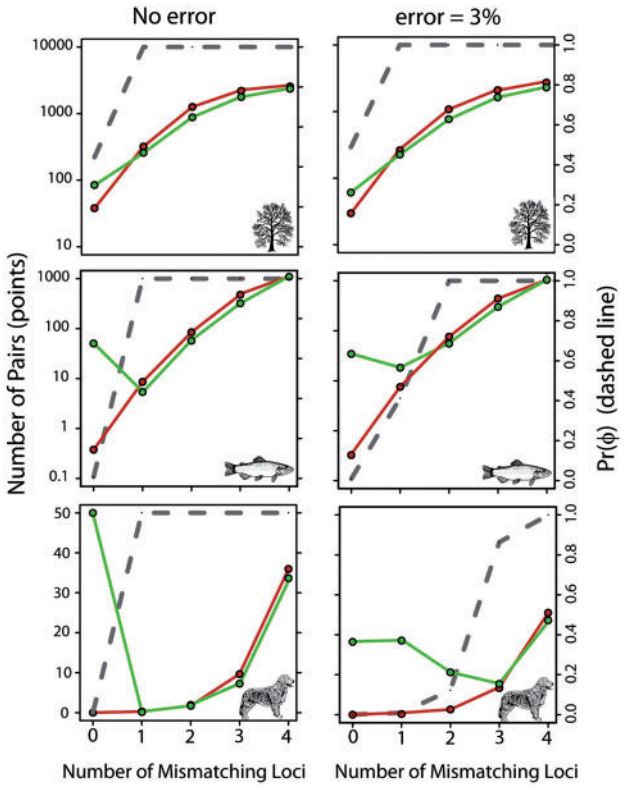

Fig. 1. Number of observed putative (*Nputative*, green points) and expected false (*Fpairs*, brown points) parent–offspring pairs in the test datasets derived from three empirical studies (Table 1). The left-hand plots represent datasets with no genotyping error, and the right-hand plots represent datasets with 3% genotyping error. Each panel represents 100 test datasets with 100 adults, 100 juveniles and 50 true parent–offspring pairs. The dashed line corresponds with the right-hand axis and represents the prior probability of a putative parent–offspring pair occurring by chance, $\Pr(\phi)$, estimated as *Fpairs*/*Nputative*. The number of true parent–offspring pairs is estimated as the difference between *Nputative* and *Fpairs*. Thus, whenever *Nputative* is greater than *Fpairs*, $\Pr(\phi)$ is <1, and a nonzero proportion of true parent–offspring pairs may be inferred

estimates of allele frequencies from programs that specialize with such data types (e.g. MICROCHECKER, van Oosterhout *et al*., 2006). To our knowledge, this is the first parentage method that can account for genotyping errors without needing estimates of the genotyping error rate.

2.3 Microsatellites versus SNPs

Using hundreds of thousands to millions of SNPs can allow for the elucidation of first, second and third order relatives (Manichaikul *et al*., 2010). Nevertheless, for most species, it is not yet cost effective to genotype hundreds or thousands of individuals at so many markers. SOLOMON cannot expediently process millions of SNPs, but it can accommodate large SNP datasets by performing *a priori* power analyses to determine a minimum number of SNPs for the given sample sizes to capture all true parent–offspring pairs. After a conservative number of SNPs are determined, the appropriate number of loci can be selected.

The precision associated with the posterior probabilities is increased by increasing the number of simulated datasets and genotypes. Because of the greater number of alleles and lower numbers of loci typically found in microsatellite studies, these markers require more simulations than SNPs for comparable levels of precision (Supplementary Table S1 for details and guidelines).

2.4 Validation

We use hypothesis-testing nomenclature to define the null hypothesis as no relationship between a putative parent–offspring pair (i.e. the pair is unrelated). In this framework, a type I error occurs when a putative pair is unrelated, but are falsely identified as a true pair for a given alpha. For example, a type I error would occur if alpha was set to 0.05 and an unrelated adult and juvenile were assigned a $\Pr(\phi|\lambda)$ value <0.05. Because lower $\Pr(\phi|\lambda)$ values represent a reduced probability of sharing alleles by chance, a lower posterior probability represents a reduced probability of committing a type I error. For most methods, the type I error should be less than or equal to the chosen alpha, else too many alternative hypotheses will be falsely accepted. A type II error occurs when a true parent–offspring pair is not identified for a given alpha (i.e. $\Pr(\phi|\lambda)>\alpha$ for a true parent–offspring relationship). We determined the properties of our method by measuring the type I and type II errors across a range of alpha levels.

To examine the relationship between alpha and type I and II errors, we used the per locus allele frequencies from the empirical studies (Table 1) to construct test datasets. For each of the three empirical studies, we created 100 test datasets with 100 adults, 100 juveniles and 50 true parent–offspring pairs. The adult and juvenile genotypes were created in accordance with HWE. The parents and offspring were created by randomly selecting 50 adults and 50 juveniles and, for each pair, randomly copying one allele from the adult to the juvenile at each locus. For each of the 100 test datasets, the posterior probabilities were calculated, and type I and type II errors were identified. Precision of the posterior probability was calculated by measuring the range of posterior probabilities across identical pairs from 100 replicate runs of a single test dataset from each of three study species (Supplementary Table S1). We also created test datasets with varied numbers of unrelated individuals and offspring per parent (Supplementary Tables S2 and S3).

We examined the effects of genotyping error by introducing errors into the test datasets. We defined the genotyping error rate as the proportion of all alleles that were called incorrectly (Bonin *et al.*, 2004; Pompanon *et al.*, 2005). To add error to the test datasets, we randomly sampled a single allelic position from the multilocus dataset. We treated the dataset as a matrix with m rows and n columns and randomly selected allele α_{mn}. We next replaced allele α_{mn} with a randomly selected allele from the same locus. This process was repeated until the desired genotyping error rate was obtained. Because alleles were randomly selected, an allele chosen to contain an error could be replaced with the same allele. We chose genotyping error rates of 0, 0.005, 0.01 and 0.03 because they encompass the average documented error rates for SNPs and microsatellites (Pompanon *et al.*, 2005; Saunders *et al.*, 2007).

2.5 Number of known parents

The approach presented earlier in the text is general in that no information about the sample of adults is required. We expanded the aforementioned approach to two specific parentage applications. First, we expanded the method to situations where one parent is known and it is possible to genotype the known parent and their offspring. For example, many young mammals remain closely associated with their mothers. After genotyping both the mother and their offspring, it is possible to exclude the maternal alleles from the offspring. This reduces the number of alleles to search for in putative fathers and can greatly increase the power for assignment (Jamieson and Taylor, 1997). Second, we expanded the approach to include known parent pairings, where it is known which males mated with which females. For example, captive breeding and livestock programmes often specifically cross certain males with females and keep detailed records of such pairings. Knowing which females and males are paired can substantially increase assignment power because it (i) reduces the number of pairwise comparisons and (ii) each allele in the offspring must match one allele in each parent. To allow researchers to take advantage of the increased power and reduced type I error from such study designs, we appropriately modified the simulation and posterior probability calculation algorithms. We tested these modified approaches with 100 test datasets created from the European beech study because it had the lowest power of the three datasets (and thus the most to gain from additional information). For validation purposes, we set the genotyping error rate to 1% and created mother–father pairs (100 mothers:100 fathers), each of which produced a single offspring.

2.6 Siblings and other relatives

Although full-siblings differ from parents and offspring in the way that alleles are shared by descent (Blouin 2003), they can share alleles across large numbers of loci, particularly when including alleles that are shared by chance. This is only a concern if full-siblings occur in both the sampled adults and juveniles (e.g. species with lengthy and overlapping generation times), and if they occur at high frequency. To account for full-siblings, we additionally calculate a modified Bayesian prior that includes alleles that are both identical-by-state and identical-by-descent. This modification results in a more conservative test that prevents full-siblings from being assigned as parent–offspring pairs. We tested both the modified and unmodified approach on datasets as described earlier in the text, but where we introduced pairs of full-siblings as 5, 15, 25 and 50% of the sampled individuals. Additionally, we tested whether more distant kinship pairs (e.g. aunts/uncles to nieces/nephews, half-siblings) would be falsely identified as parent–offspring pairs.

2.7 Comparison with existing methods

We next analysed empirical data by examining paternity assignments for 4 run-years of summer-run steelhead collected from the Hood River, OR, USA. This is a new dataset that has not been previously analysed. Tissue samples from all returning anadromous steelhead were collected, as the fish were passed

over the Powerdale dam en route to their spawning grounds. The dam was a complete barrier to migrating fish. All 1702 summer-run steelhead were genotyped at the same eight polymorphic loci used in the winter-run steelhead examples previously mentioned (Araki *et al*., 2007). This dataset presents a rigorous test for two reasons. First, not all candidate fathers were sampled because resident steelhead (i.e. rainbow trout) that remained above the dam could also have sired offspring (Christie *et al*., 2011). Second, any given offspring may have aunts and uncles competing for parentage assignments (Olsen *et al*., 2001).

Direct and equitable comparisons between parentage methods can be challenging because each method represents different theoretical approaches. Furthermore, each method often makes different assumptions and requires different input information. We first used Mendelian incompatibility (exclusion) to assign offspring to putative fathers. We allowed one locus to mismatch to account for genotyping error. We next employed the most-frequently used parentage program, CERVUS 3.03 (Kalinowski *et al*., 2007; Marshall *et al*., 1998), to perform the same assignments. CERVUS uses a simulation procedure to determine the significance of log-likelihood scores for candidate parent–offspring pairs. This program requires the estimates of three parameters: (i) the number of candidate parents, (ii) the proportion of candidate parents sampled and (iii) the genotyping error rate. Because we did not have estimates of these parameters (they require substantial observational data), we set the number of candidate parents to the number of adults sampled in our dataset and chose a small and large proportion of candidate parents sampled (0.1 and 0.9, respectively). We set our genotyping error rate to 1%, which is the default setting, and included assignments with 95% or higher confidence. Finally, we used SOLOMON to analyse the same sets of samples, using an alpha of 0.05.

To verify our assignments with these three methods, we genotyped all individuals at five additional microsatellite loci (see Supplementary Material for details). To determine which pairs were definitively true, we performed exclusion at all 13 loci and allowed for one locus to mismatch. For matches at both 12 and 13 loci, the average expected number of false pairs was <1. For all three methods, we measured the total number of assignments and the total number of correct assignments as determined by comparison with the pairs identified with the additional loci.

3 RESULTS

3.1 Validation

For all three empirical studies used to generate test datasets, the type I error rate was always equal to or less than the desired alpha (Fig. 2). The beech datasets had the highest type II error rate (lowest power) of the three studies. The steelhead datasets had a lower type II error rate, despite having five fewer loci than the beech study. Thus, in these two cases, increased marker polymorphism resulted in greater power for parentage analysis than did additional loci. Finally, the retriever study with 200 SNPs had the lowest type II error rate (highest power), further confirming that SNPs can be useful markers for parentage analysis (Anderson and Garza, 2006). The inherent tradeoffs between type I and II errors revealed that there is a marked decrease in

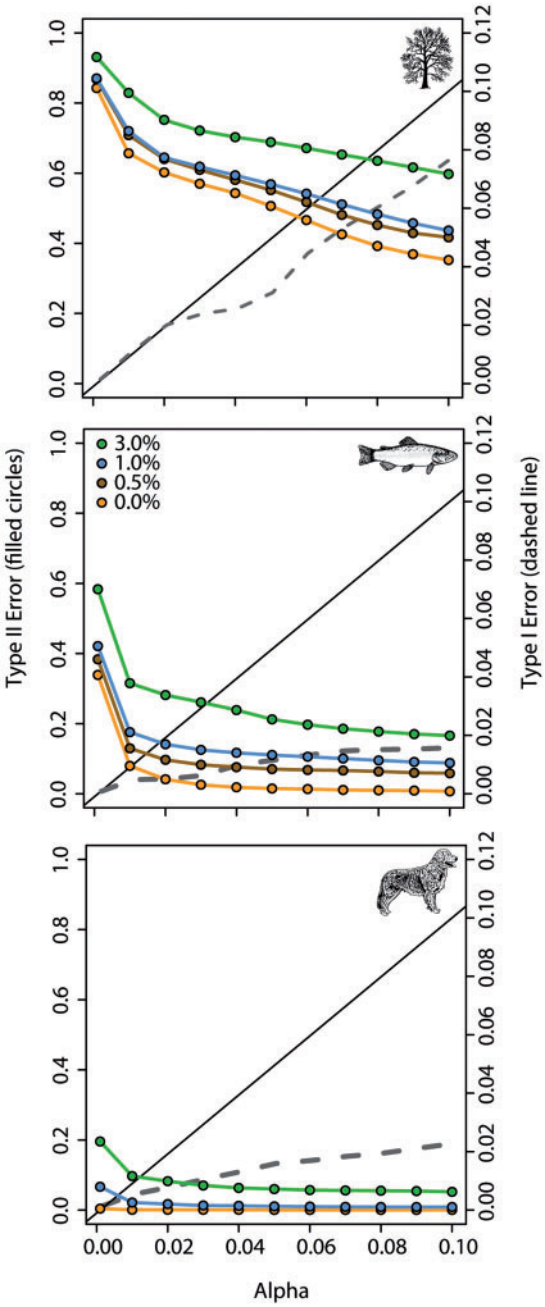

Fig. 2. The relationship between alpha and the type I and II error rate. Genotyping error rates were varied from 0 to 3% (line order corresponds to legend order). Each panel represents 100 test datasets with 100 adults, 100 juveniles and 50 true parent–offspring pairs. The maximum observed type I error was plotted as a dashed grey line. Type I error is consistently at or below α (solid line), indicating that our method is conservative and does not produce an excess of false-positive parent–offspring pairs. For the steelhead and Labrador retriever datasets, an increase in alpha beyond 0.05 recovers few additional true parent–offspring pairs. The lowest alpha value plotted is 0.001, and the 0.5% genotyping error was omitted from the retriever dataset for visual clarity. See Supplementary Figure S1 to view these results on a logarithmic scale

type II error (increase in power) by changing the alpha threshold from 0.001 to 0.01. Further increases in alpha from 0.01 to 0.1 yielded marginal increases in power for the steelhead and retriever datasets, but it provided consistent increases in power for the beech dataset. In general, a good tradeoff between type I and II errors can be obtained by setting alpha at 0.05, but this value should ultimately be decided by weighing the relative risks of committing type I and II errors for a particular study (Sokal and Rohlf, 1994). Not surprisingly, the likelihood of committing type I errors increases with low-power datasets that have high values for the prior. As such, we recommend reporting both the prior and posterior probabilities.

In all three datasets, genotyping error increased the number of type II errors. Because the retriever dataset could allow for the greatest number of mismatching loci (Fig. 1), this dataset was the least affected by genotyping error. In general, genotyping error rates of 0.005 or 0.01 did not drastically increase the type II error rate. A genotyping error rate of 3%, however, did result in substantial increases in type II error for all three datasets. We further examined the tradeoff between genotyping error rates and power in the retriever dataset. All datasets, regardless of the genotyping error rate, identified all true parent–offspring pairs with 250 loci (Fig. 3). As expected, the number of loci required to identify all true parent–offspring pairs increased with an increase in the genotyping error rate.

Additional samples of a single known-parent or information about putative parent pairings greatly reduced the type I and II error rates (Fig. 4). Both the type I and type II errors were highest when no known parents were sampled. Having a known sample of one of the parents or knowing the parent pairs reduced the type II error by nearly 60% for the beech study. Thus, when possible, we recommend collecting this additional data to maximize power for parentage analysis.

In general, pairs of simulated full-siblings that were split between adult and juvenile files did not get assigned in large numbers until they represented >25% of the individuals in a dataset (Supplementary Table S4). Adjusting the prior for alleles that were identical-by-state as well as those that were identical-by-descent resulted in fewer sibling pairs identified with a posterior probability <0.05 (Supplementary Table S5). Accounting for alleles that are identical-by-descent comes at the cost of assigning true parents, as it can be difficult to distinguish between full-siblings and parent–offspring pairs with genotyping errors with limited numbers of loci. As such, we recommend using the modified sibling approach only when large numbers of siblings are expected to be sampled. Other levels of relationship that share fewer alleles than full-siblings (e.g. aunts/uncles to nieces/nephews) were not falsely identified using the unmodified approach.

3.2 Empirical data

Across all 4 run-years of our summer-run steelhead dataset, we found that using simple exclusion for seven of eight loci (i.e. allowing one locus to mismatch) resulted in a high type I error rate. Using exclusion, 349 offspring were assigned to a father, of which 213 were later confirmed to be true assignments with genotyping at the five additional loci (Table 2). Thus, exclusion produced 136 false assignments, yielding a type I error rate of 0.39. CERVUS had type I error rates of 0.22 and 0.49 when we set the estimates of the proportion of candidate parents sampled to 0.1 and 0.9, respectively. In contrast, SOLOMON had a type I error rate of 0.029 for an alpha set to 0.05. Consistent with the results from the test datasets (Figs 2 and 4), varying the alpha in this empirical dataset resulted in an

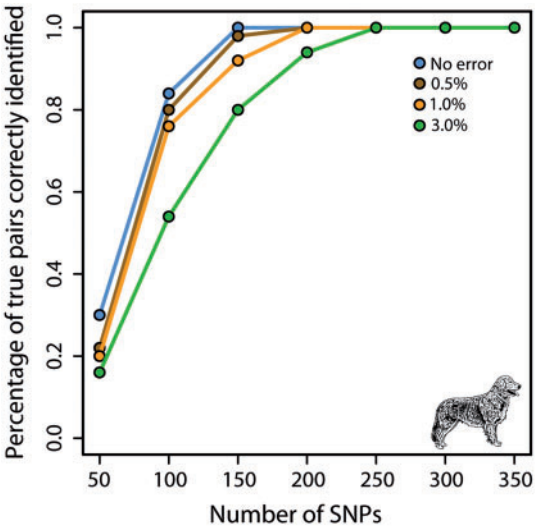

Fig. 3. Relationship between the number of used SNPs and the percentage of true parent–offspring pairs that were correctly identified in the retriever datasets. Genotyping error rates were varied from 0 to 3%, and all parent–offspring pairs were correctly identified with 250 SNPs (line order corresponds to legend order). Notice that small amounts of error do not substantially affect the assignment rate with intermediate numbers of loci

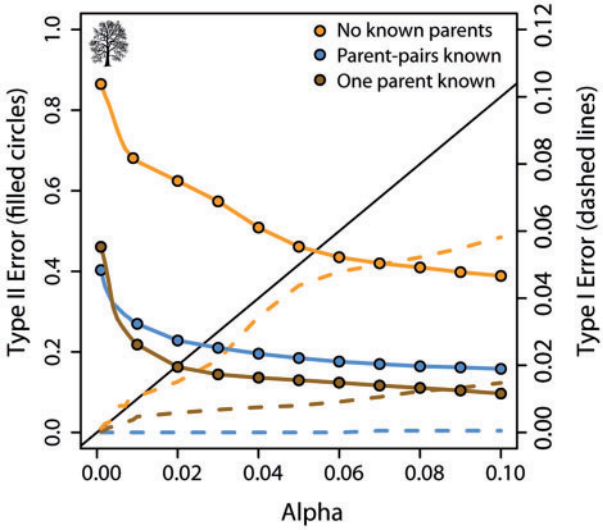

Fig. 4. The relationship between alpha and the type I and II error rate for three parentage scenarios: no known parents (orange circles), known parent pairs (blue circles) and one known parent (brown circles). Notice that type I and II errors are reduced, as additional parentage information is used. For each parentage scenario, 100 test datasets were constructed with 100 adults, 100 juveniles and 100 true parent–offspring pairs

Table 2. Comparison of exclusion, CERVUS and SOLOMON on a summer-run steelhead dataset

Run-year	Adults/ juveniles	Method	Assigned	Correct
2001	201/227	Exclusion	79	38
2001	201/227	CERVUS	35 (98)	23 (37)
2001	201/227	SOLOMON	29	27
2002	343/285	Exclusion	141	90
2002	343/285	CERVUS	47 (151)	39 (78)
2002	343/285	SOLOMON	63	61
2003	144/216	Exclusion	73	49
2003	144/216	CERVUS	44 (83)	34 (49)
2003	144/216	SOLOMON	28	28
2004	90/196	Exclusion	56	36
2004	90/196	CERVUS	32 (65)	27 (35)
2004	90/196	SOLOMON	20	20
All years	778/924	Exclusion	349	213
All years	778/924	CERVUS	158 (397)	123 (199)
All years	778/924	SOLOMON	140	136

Adults/juveniles column represents the sample sizes of adults and their putative offspring, respectively. Assigned refers to the total number of assignments. Correct refers to the number of assignments that were correct after genotyping all putative pairs at five additional loci. For CERVUS, we estimated the proportion of candidate parents sampled to be 0.1 or 0.9, although we did not possess demographic estimates of this parameter (results for 0.9 are presented in parentheses). Using more moderate values for the proportion of candidate parents sampled (e.g. 0.05) did not improve performance.

observed type I error less than or equal to alpha in all 4 years (Supplementary Table S6). It is worth noting that in some years, CERVUS had a higher number of false assignments than exclusion because the program sometimes allowed for up to two loci to mismatch.

Previous studies have shown that the performance of CERVUS is robust, and we suspect that the possible presence of aunts and uncles among the candidate parents coupled with unknown numbers of sampled parents provided challenging conditions. In general, SOLOMON performed favourably by minimizing the number of false assignments while maximizing the number of correct assignments (Table 2).

4 DISCUSSION

Accurate parentage assignments are necessary to appropriately address a wide range of research questions (Jones and Ardren, 2003; Pemberton, 2008). Here, we provide a Bayesian method that can account for genotyping error, missing data and false matches without requiring estimates of any non-genetic parameters (i.e. all analyses simply use the provided genotypic data). These methods can be applied to a broad array of datasets ranging from samples of large wild populations with unknown numbers of sampled parents to carefully controlled crosses with detailed pedigree records. To our knowledge, this is the first parentage program that does not require direct estimates of genotyping error. This solution represents a significant advance because choosing the appropriate method for estimating genotyping error rates can be ambiguous and is further obfuscated by the different types of genotyping errors that can occur (Pompanon *et al.*, 2005). Furthermore, the estimation of error rates typically involves the genotyping of additional (or duplicate) samples, which is costly from both a time and monetary standpoint. Because this method was designed with a null hypothesis of no relationship, it may not be ideally suited for datasets with large numbers of related individuals (e.g. full-sibs). Future improvements could include specifying different null hypotheses of relationship and evaluating them in a likelihood-based framework.

Our analyses revealed that, for a given dataset, the Bayesian approach appropriately minimizes false assignments while maximizing the number of correct assignments. The number of true parent–offspring relationships correctly identified depends on the sample sizes, the number of loci, the allele frequencies and the genotyping error rate. For a given marker set, larger sample sizes rapidly increase the number of pairs that share alleles by chance (Christie 2010), and increases in genotyping error can diminish power (Figs 2 and 3). Furthermore, the number and frequency distribution of alleles at each locus contribute to the rate of false matching. Uniform allele frequencies result in the greatest power for parentage analysis, but they are rarely observed in genetic markers. On the other hand, SNPs with a minor allele frequency <1% will contribute little information to the elucidation of parent–offspring pairs. Given the multitude of factors that contribute to false matching and reduced power, we suggest that researchers conduct *a priori* power analyses before designing a study that involves parentage analysis. Such power analyses can dictate precisely how many loci would be required for given sample sizes.

We provide a module for *a priori* power analysis as part of our program SOLOMON, which is available as a freely distributable R package (R Development Core Team, 2012) from the CRAN website: http://cran.r-project.org/web/packages/. SOLOMON is run with a graphical user interface written with the TCL/TK package provided by R. SOLOMON performs the described Bayesian parentage analysis for datasets with no known parents, one known parent or known parent pairs. Using an Intel core i7^{TM} processor with 8 GB of random access memory, the average run-time was 11 min for the beech datasets, 8 min for the steelhead dataset and 13 min for the retriever dataset (with larger sample sizes resulting in increased run times). Furthermore, the program performs exclusion for the three types of parentage analysis, and the exclusion interfaces allow for user-defined numbers of loci to mismatch. In summary, the Bayesian approach implemented in SOLOMON can be applied to a wide variety of datasets, resulting in robust parentage assignment.

ACKNOWLEDGEMENTS

The authors acknowledge Zaid Abdo, Chris Sullivan and the Center for Genome Research and Biocomputing at Oregon State University for helpful contributions. They also thank the reviewers for comments that greatly benefited this manuscript.

Funding: Bonneville Power Administration (to M.S.B.).

Conflict of Interest: none declared.

REFERENCES

Akey,J.M. *et al.* (2010) Tracking footprints of artificial selection in the dog genome. *Proc. Natl. Acad. Sci. USA*, **107**, 1160–1165.

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

Araki,H. *et al.* (2007) Reproductive success of captive-bred steelhead trout in the wild: evaluation of three hatchery programs in the Hood river. *Conserv. Biol.*, **21**, 181–190.

Blouin,M.S. (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol. Evol.*, **18**, 503–511.

Bonin,A. *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Mol. Ecol.*, **13**, 3261–3273.

Christie,M.R. (2010) Parentage in natural populations: novel methods to detect parent-offspring pairs in large data sets. *Mol. Ecol. Resour.*, **10**, 115–128.

Christie,M.R. *et al.* (2011) Who are the missing parents? Grandparentage analysis identifies multiple sources of gene flow into a wild population. *Mol. Ecol.*, **20**, 1263–1276.

Hadfield,J.D. *et al.* (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Mol. Ecol.*, **15**, 3715–3730.

Jamieson,A. and Taylor,S.S. (1997) Comparisons of three probability formulae for parentage exclusion. *Anim. Genet.*, **28**, 397–400.

Jones,A.G. and Ardren,W.R. (2003) Methods of parentage analysis in natural populations. *Mol. Ecol.*, **12**, 2511–2523.

Jones,A.G. *et al.* (2010) A practical guide to methods of parentage analysis. *Mol. Ecol. Resour.*, **10**, 6–30.

Kalinowski,S.T. *et al.* (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.*, **16**, 1099–1106.

Lander,T.A. *et al.* (2011) Reconstruction of a beech population bottleneck using archival demographic information and Bayesian analysis of genetic data. *Mol. Ecol.*, **20**, 5182–5196.

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

Marshall,T.C. *et al.* (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.*, **7**, 639–655.

Olsen,J.B. *et al.* (2001) The aunt and uncle effect: an empirical evaluation of the confounding influence of full sibs of parents on pedigree reconstruction. *J. Hered.*, **92**, 243–247.

Pemberton,J.M. (2008) Wild pedigrees: the way forward. *Proc. Biol. Sci.*, **275**, 613–621.

Pompanon,F. *et al.* (2005) Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.*, **6**, 847–859.

Rieseberg,L. *et al.* (2012) Editorial 2012. *Mol. Ecol.*, **21**, 1–22.

Saunders,I.W. *et al.* (2007) Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics*, **90**, 291–296.

Slate,J. *et al.* (2000) A retrospective assessment of the accuracy of the paternity inference program CERVUS. *Mol. Ecol.*, **9**, 801–808.

Sokal,R.R. and Rohlf,F.J. (1994) *Biometry*, 3rd edn. W.H. Freeman, New York, USA.

van Oosterhout,C. *et al.* (2006) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Mol. Ecol. Notes*, **6**, 255–256.

Vandeputte,M. *et al.* (2006) An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. *Mol. Ecol. Notes*, **6**, 265–267.