Received: 7 August 2017 | Revised: 16 February 2018 | Accepted: 24 February 2018

DOI: 10.1111/1755-0998.12782

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Efficient inference of paternity and sibship inference given known maternity via hierarchical clustering

Thomas James Ellis[1,2] | David Luke Field[1,3] | Nicholas H. Barton[1]

[1]Institute of Science and Technology Austria, Klosterneuburg, Austria

[2]Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[3]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

Correspondence
Nicholas H. Barton, Institute of Science and Technology Austria, Klosterneuburg, Austria.
Email: nick.barton@ist.ac.at

Funding information
ERC, Grant/Award Number: 250152

Abstract

Pedigree and sibship reconstruction are important methods in quantifying relationships and fitness of individuals in natural populations. Current methods employ a Markov chain-based algorithm to explore plausible possible pedigrees iteratively. This provides accurate results, but is time-consuming. Here, we develop a method to infer sibship and paternity relationships from half-sibling arrays of known maternity using hierarchical clustering. Given 50 or more unlinked SNP markers and empirically derived error rates, the method performs as well as the widely used package *Colony*, but is faster by two orders of magnitude. Using simulations, we show that the method performs well across contrasting mating scenarios, even when samples are large. We then apply the method to open-pollinated arrays of the snapdragon *Antirrhinum majus* and find evidence for a high degree of multiple mating. Although we focus on diploid SNP data, the method does not depend on marker type and as such has broad applications in nonmodel systems.

KEYWORDS

Antirrhinum, fractional assignment, paternity, pedigree, sibships

1 | INTRODUCTION

The genealogical pedigree of a population is a valuable piece of information for genetic studies because it conveys complete information about the relatedness between individuals (Pemberton, 2008). The pedigree of wild populations can be inferred from patterns of shared alleles among individuals based on marker data (Blouin, 2003; Jones, Small, Paczolt, & Ratterman, 2010). The simplest approach is to infer the parents of individual offspring and build a pedigree from pairs or triplets of individuals (Marshall, Slate, Kruuk, & Pemberton, 1998; Meagher, 1986; Meagher & Thompson, 1986). However, the accuracy of pedigree reconstruction can be increased by including as many sources of relevant information as possible (Neff, Repka, & Gross, 2001; Wang, 2007). One way to achieve this is to reconstruct parentage and sibling relationships simultaneously, because alleles shared among full-siblings allow us to better identify their common parents that would be possible by considering each offspring individually (Sieberts, Wijsman, & Thompson, 2002; Wang & Santure, 2009). A major challenge is to find the best way to partition offspring groups to distinguish full-, half- and nonsiblings, because the number of possible configurations is too large to enumerate, even for modest family sizes.

Given the complexity of this problem, current methods employ an iterative search via a Markov chain algorithm (e.g., Anderson & Ng, 2016; Emery, Wilson, Craig, Boyle, & Noble, 2001; Jones et al., 2007; Thomas & Hill, 2002; Wang, 2004, 2012; Wang & Santure, 2009). For example, Colony repeatedly merges and splits full-sibship groups and infers the likely genotype of the parents of each group by simulated annealing (Wang, 2004; Wang & Santure, 2009). This provides accurate solutions, but can be slow, especially for large data sets, which limits their applicability to larger data sets or more complex downstream analyses. Wang (2012) altered the method to run more efficiently with SNP data. This was found to be less accurate than the full-likelihood method, but provides a substantial increase in speed and computational burden. In spite of this boost, it still relies on a Markov chain algorithm and remains time-consuming. As such, there is still scope to improve the efficiency method to infer sibling and paternity relationships from genetic markers.

An alternative way to group individuals into sibship groups is through hierarchical clustering based on a metric of relatedness. Hierarchical clustering is a flexible machine-learning technique to identify plausible ways to group items based on some measure of how similar (or dissimilar) each pair of items are to one another (Murtagh & Contreras, 2012). First, a matrix of similarity between all pairs of items is created, and the closest items are joined sequentially to form distinct groups. This grouping is based on a linkage function of the distance between groups. Common linkage functions include the minimum, maximum or mean distance, but other functions are common. Hierarchical clustering does not return a single optimal configuration of groups, but rather a hierarchy of possible configurations that can be compared. This provides a means to quickly generate samples of plausible sibship configurations and to assess the relative support for each.

In the absence of perfect data, inference from pedigrees will be more reliable if we also account for uncertainty about parentage. In the simplest (and probably most commonly applied) approach, categorical parentage methods identify the most likely pedigree, which is assumed to be correct in subsequent analyses (e.g., Anderson & Garza, 2006; Jones & Ardren, 2003; Marshall et al., 1998; Meagher, 1986). This assumption will tend to bias paternity towards highly homozygous candidates (Devlin, Roeder, & Ellstrand, 1988). Markov chain algorithms can account for uncertainty in parentage structure by sampling likely candidates in turn and provide a posterior distribution of possible pedigrees (Anderson & Ng, 2016; Emery et al., 2001; Hadfield, Richardson, & Burke, 2006; Jones et al., 2007; Wang & Santure, 2009). An alternative, but overlooked, framework for parentage inference is fractional parentage assignment, where all candidate parents are assigned a probability of parentage for each offspring (Devlin et al., 1988; Nielsen, Mattila, Clapham, & Palsbøll, 2001; Roeder, Devlin, & Lindsay, 1989). This can be expressed efficiently in matrix form and allows us to consider the probability of parentage for all candidates simultaneously. However, although this is valid for the parentage of individual offspring, it is more challenging to deal with the parentage of full-sibships in a fractional framework, because, by definition, no two families can share both parents.

In this study, we outline methods to jointly infer sibling and paternal relationships using hierarchical clustering within a fractional framework. To simplify presentation, we focus on inference of paternity, where the identity of the maternal parent is known with certainty, although in principle the method is equally valid if both parents are unknown. We present the Python package fractional analysis of paternity and sibships (FAPS) that allows users to implement the method and easily draw inferences about family structure that automatically accounts for uncertainty about paternity and sibling relationships. Using simulations, we demonstrate that the method is robust to genotype quality, sample size and mating patterns. We then apply the method to wild seedlings of the snapdragon *Antirrhinum majus* where there are a very large number of candidate males and find evidence for high degree of polyandry.

2 | METHODS

2.1 | Paternity of individuals

In this study, we consider the case of half-sibling arrays, where a set of offspring individuals $O = \{o_1, o_2, \ldots, o_{n_o}\}$ sharing a single mother m are arranged into one or more full-sibships. Assuming for the moment that all males in the population have been sampled, each offspring individual and each full-sibship have a single father in the set of candidate fathers $F = \{f_1, f_2, \ldots, f_{n_f}\}$.

Our method relies on individual-paternity matrix **G**, describing all possible pairwise relationships between offspring and candidate males. Each row of **G** corresponds to a single offspring individual and each column to a candidate father. Element g_{ij} of **G** is proportional to the likelihood that the jth candidate father in F is the true father of the ith offspring in O, based on the genotype data for candidate, offspring and mother (Thompson & Meagher, 1987; Appendix 1). Each row of **G** can be seen as a probability distribution of possible paternities for a single offspring, and as such, each row of **G** must sum to one (Devlin et al., 1988; Nielsen et al., 2001). Given complete sampling of males and uniform prior belief about the importance of each male, g_{ij} is simply

$$g_{ij} = \frac{L(o_i|f_j, m)}{\sum_f L(o_i|f_j, m)} \tag{1}$$

where $L(o_i|f_j, m)$ is the likelihood of generating the observed offspring genotype given the genotypes of m and the jth candidate male. If sampling is not complete, **G** should be adjusted accordingly (Nielsen et al., 2001; Appendix 1).

Methods to calculate $L(o_i|f_j, m)$ from marker data are well established (Thompson & Meagher, 1987; Meagher & Thompson, 1986; Marshall et al., 1998; see Jones et al., 2010 for a practical review). It is important that **G** should account for errors and missing data in genotype information, and the most appropriate method for doing this will depend on the kind of marker being used, such as microsatellites (Marshall et al., 1998; Wang, 2004) or SNPs (Anderson & Garza, 2006). Unless otherwise stated, we have followed Anderson & Garza (2006) and Nielsen et al. (2001) in the calculation of likelihoods of paternity for sampled and missing candidates, respectively (Appendix 1).

2.2 | Clustering into full-sibships

Let T_c be a set of full-sibship groups that partition the n offspring in half-sibling array O into between one and n subsets of full-siblings. The kth full-sibship group in T_c is $t_k \subseteq T_c$. In a fractional framework, we aim to identify a set of likely configurations $\{T_1, T_2, \ldots, T_n\}$ and account for the relative probability of each.

We use the Unweighted Pair Group Method with Arithmetic Mean Algorithm (UPGMA) to cluster individuals into full-sibship families (Sokal & Michener, 1958). This is a specific instance of hierarchical clustering algorithms that use the means distance between groups as its linkage function. We first calculate distance matrix **D**, whose *ih*th element is based on the likelihood $\sum_j g_{ij}g_{hj} = \boldsymbol{g_i} \cdot \boldsymbol{g_h}$ that the *i*th and *h*th offspring are full-siblings, where *j* indexes each father in *F*. This is converted to a distance metric by taking the absolute value of the log of this value:

$$\mathbf{D_{ih}} = \left| \ln \sum_j g_{ij}g_{hj} \right|. \tag{2}$$

We then use the UPGMA to build a dendrogram of individuals based on **D**. By bisecting this dendrogram at different heights, the *n* offspring can be grouped into *n* unique configurations (one for every internal node of the dendrogram and one for the tips; see Figure 1). This approach provides a sample of likely partitions without the need to explore the sample space of possible partitions iteratively.

2.3 | Likelihood of a partition

To qualify as a full-sibship, all individuals in putative subset t_k must share the same father. For a given partition structure, the likelihood that the *j*th father is the true father of every individual in t_k is proportional to the product of his probability of paternity for each offspring, $\prod_{i \in t_k} g_{ij}$, and the likelihood for t_k is the sum over likelihoods for each father, $\sum_j \prod_{i \in t_k} g_{ij}$. If each sibship were independent, the likelihood of the proposed partition would simply be the product of the likelihoods for each set of full-siblings:

$$\Pr(T_c) \propto \prod_k \sum_j \prod_{i \in t_k} g_{ij}, \tag{3}$$

where *j* indexes candidate fathers, *i* indexes offspring, and *k* indexes full-sibships.

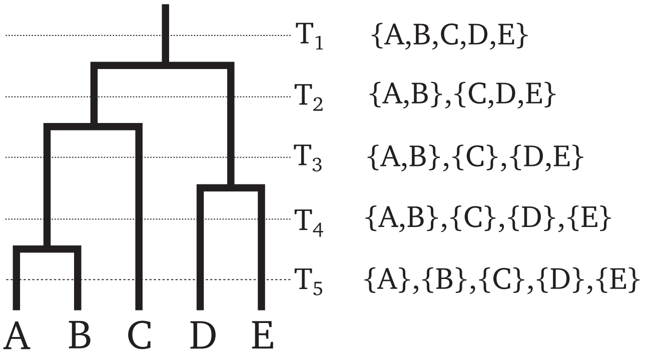

FIGURE 1 Partitioning into full-sibships using a dendrogram. A dendrogram can be constructed from a matrix of relatedness for five (A–E) individuals in a half-sibling array. By bisecting the dendrogram at different positions, there are five unique partitions to group individuals into possible full-sibships, labelled T_1–T_5. Partition structures are shown on the right

However, to be distinct families, each subset t_k may not share its father with any other subset $t_{m \neq k}$ in T_c. This means that Equation 3 is not valid unless the set of compatible fathers for each full-sibship is unique, with no intersection with that of any other full-sibship. For example, imagine two proposed full-sibships, *a* and *b*, and three candidate fathers with nonzero probabilities of paternity for each full-sibship. Then, $\gamma_a = \{a_1, a_2, a_3\}$ and $\gamma_b = \{b_1, b_2, b_3\}$, where the subscript indexes the likelihood of paternity for each candidate male on each full-sibship. Equation 3 is then

$$\sum_j \gamma_{aj} \sum_j \gamma_{bj} = (a_1 + a_2 + a_3)(b_1 + b_2 + b_3)$$

which expands to a sum of nine terms for all possible pairwise combinations:

$$\begin{matrix} a_1b_1 & a_1b_2 & a_1b_3 \\ a_2b_1 & a_2b_2 & a_2b_3 \\ a_3b_1 & a_3b_2 & a_3b_3 \end{matrix}$$

To account for terms describing shared fathers across sibships, it is necessary to remove those terms with one or more matching subscripts, which in this example are the diagonal elements. For small cases such as this, it is straightforward to identify these terms and subtract them from Equation 3, but this quickly becomes computationally demanding as the number of sibships and candidate fathers increases. For this reason, we should expect Equation 3 to be intractable for most real-world cases.

Nevertheless, $\Pr(T_c|\mathbf{G})$ can be efficiently estimated using Monte Carlo simulations. Let Γ denote the matrix of probabilities of paternity for each sibship given partition structure T_c. Like **G**, Γ has one column for every candidate father, but rows correspond to proposed full-sibships rather than individuals. Element γ_{kj} of Γ is the probability that the *j*th father is the true father of every individual in t_k and is proportional to the product of his probability of paternity for each offspring in t_k:

$$\gamma_{kj} = \frac{\prod_{i \in t_k} g_{ij}}{\sum_j \prod_{i \in t_k} g_{ij}} \tag{4}$$

where *i* indexes offspring. As in Equation 1, the numerator ensures that the probabilities of paternity for all candidates sum to one. Because each sibship has exactly one father, we can sample a valid set of possible fathers for each sibship by traversing a path through Γ from top to bottom, visiting each column at most once, and each row exactly once. As before, there will be generally be too many paths to enumerate. However, because most rows will contain one or a handful of large values and many small values, only a small number of the possible paths will explain most of the total probability, and we can approximate the likelihood of the partition by only considering these.

For each row in Γ, we draw a sample of likely fathers in proportion to their probability of paternity and bind these vectors into a matrix with as many rows as in Γ. We remove duplicate columns and any columns that contain the same candidate twice. The remaining columns represent sets of unique paths through Γ, drawn in proportion to the probability of those paths. The likelihood of each path through Γ is the product of the probabilities in each column, and the likelihood for the whole partition is the sum of likelihoods for all valid paths through Γ.

We note that Equation 4 is not technically correct when genotype errors exist in f_j, because offspring in t_k are no longer independent. This will deflate γ_{kj} in proportion to the size of t_k as the error is replicated over multiple offspring. Fortunately, with more offspring in a full-sibship, it is easier to identify the true father (Sieberts et al., 2002; Wang, 2007), which should more than make up for the effect of genotype error in the father, assuming genotype error rates are not very high.

2.4 | Inference of mating patterns

We can use the methods described above to inform questions about underlying biological processes. For questions about family structure and size, it is straightforward to account for uncertainty in family structure by weighting outcomes by the relative support for each partition. For example, the probability that any pair of individuals are full-siblings is the sum of the probabilities of each partition for which the individuals are full-siblings. The probability that there are x full-sibships in a half-sibling array is the sum of each $\Pr(T_c)$ for which $|T_c| = x$. Similarly, the posterior distribution of full-sibship sizes is the sum of the distributions for each T_c weighted by $\Pr(T_c)$.

If the ith offspring is assigned to full-sibship t_k in partition structure T_c, the probability that the jth candidate is the father of o_i is γ_{kj}, conditioned on T_c. The marginal probability that the jth candidate is the father of the ith offspring is then the product of each γ_{kj} in each partition structure:

$$\Pr(o_i|f_j, m, i \in t_k) = \prod_c \gamma_{kj|T_c} \Pr(T_c). \tag{5}$$

We can then estimate the distribution of fertilities among candidates by apportioning paternity of the whole half-sibship array to each candidate. The fertility of the jth father on the mother is the proportion of the n_o total offspring sired by f_j:

$$\frac{1}{n_o} \sum_i \Pr(o_i|f_j, m, i \in t_k). \tag{6}$$

This is similar to Roeder et al.'s (1989) equation for paternal fertility and would be appropriate for investigating postmating fertility processes such as pollen competition.

However, this equation is only valid when there is little sibship structure in the sample because the assumption that offspring are independent is violated. For most biological questions, it is more appropriate compare the *fecundities* of candidate fathers on the mothers by identifying independent mating events. To do this, we employ a Monte Carlo sampling scheme similar to that described above, which allows us to account for the relative probabilities of each partition structure, while avoiding the possibility of drawing the same candidate for multiple full-sibships. For partition T_c, we first draw N sets of candidate fathers for each full-sibship in proportion to their probabilities of paternity, where N is very large (≥1,000). We then remove sets for which the same candidate is represented more than once. These sets are resampled $\Pr(T_c)N$ times to weight samples by the probability of the partition. This procedure is repeated for each partition. Because $\sum_c \Pr(T_c) = 1$, this leaves a total of N sets in superset X of likely candidate fathers. The fecundity of the jth candidate on the mother is then his frequency in all sets in X.

2.5 | FAPS package

Our method is implemented in the Python package FAPS. The packaged is based on *NumPy*, with additional tools from the *fastcluster* and *pandas* packages (McKinney, 2010; Müllner, 2013; van der Walt, Colbert, & Varoquaux, 2011). The package includes functions for calculating **G** with a focus on SNP markers (Appendix 1), clustering individuals into sibships for multiple half-sibling arrays. FAPS also allows for inference of family structure and mating patterns with summary outputs that automatically account for uncertainty in genealogy. Extensive simulation tools are provided to allow power analyses for hypothetical data sets, as well as facilities for inspecting data prior to analysis. FAPS can handle multiple half-sibling arrays and makes no requirement that a candidate father sires offspring with multiple mothers. The package and accompanying tutorial are available from www.github.com/ellisztamas/faps.

2.6 | Simulations

We used FAPS' power analysis tools to investigate how well FAPS was able to infer sibship and mating patterns as the quality of genotype information, analysis parameters and underlying family structure varied. In all simulations below, we ran 300 replicates for each parameter set.

2.6.1 | Accuracy of relationship inference

We used five metrics to assess sibship and paternity assignment:

1. $p_{\text{partition}}$: the probability that the true partition structure is included in the sample of possible partitions;
2. p_{full}: the mean posterior probability that a pair of true full-siblings are inferred to be full-siblings;
3. p_{half}: the mean posterior probability that a pair of true half-siblings are inferred to be half-siblings;
4. p_{sire}: the mean posterior probability that an individual's true sire is inferred to be the true sire;
5. p_{absent}: the mean posterior probability that the true sire is inferred to be missing from the sample of candidates;
6. The distribution of inferred family sizes, or else the distribution of number of families.

With the exception of $p_{\text{partition}}$, probabilities are calculated integrating over possible partition structures.

2.6.2 | Family structure scenarios

In a first set of simulations, we investigated inference of family structure for half-sibling arrays of 20 offspring under four contrasting scenarios: (i) even sibship sizes (four full-sibships of five offspring); (ii) a single family (one full-sibship of 20 offspring); (iii) all half-siblings (20 full-sibships of one offspring); (iv) reproductive skew (one full-sibship of 10 individuals plus 10 families of one offspring).

For each scenario below, we simulated genotypes for a single mother, as well as 100, 250, 500, 1,000 or 2,000 candidate males based on between 30 and 100 unlinked SNP loci. SNP minor allele frequencies were drawn from a uniform distribution between 0.3 and 0.5. We simulated offspring genotypes based on parental genotypes and Mendelian inheritance. We then added point mutations to adult and offspring genotypes at random with per-locus probabilities of $\mu = 0.0015$, 0.005, 0.01 or 0.015 to simulate errors in genotyping.

2.6.3 | Full-sibship size

To test the effect of overall offspring sample size, we repeated simulations for the even-sibship-size mating scenario, but using full-sibship sizes of 2, 10, 25, 50 or 100 individuals. As these simulations were computationally intensive, we only used 250 candidate males, 50 loci and $\mu = 0.0015$.

2.6.4 | Number of Monte Carlo draws

To investigate the sensitivity of sibship inference to the number of Monte Carlo draws used to estimate the likelihood of a partition structure, we repeated simulations of the four family structure scenarios described above using 10^1, 10^2, 10^3 and 10^4 draws. We performed simulations assuming $\mu = 0.0015$ and 50 loci. For each simulated data set, we assessed how much of the probability space of possible sibship and paternity relationships had been explored by summing the likelihoods for each partition structure inferred from hierarchical clustering. We also inferred the effect of changing the number of draws on the accuracy of sibship inference.

2.6.5 | Comparison with Colony

We compared the performance of FAPS with the serial command-line version of Colony 2.0.6.3 for Linux (Wang & Santure, 2009). We used FAPS to simulate half-sibling arrays containing four full-sibships of five individuals and 250 candidate males, using between 10 and 80 loci and $\mu = 0.0015$. We then used FAPS and Colony to infer family structures of each half-sibling array assuming no self-fertilization.

Colony allows for three analysis methods: (i) "full likelihood" (FL) that incorporates information on both sibship and parental relationships jointly (Wang, 2004); (ii) "pairwise likelihood" (PLS) that reconstructs relationships for pairs of individuals only; and (iii) a hybrid method which is designed to work efficiently with SNP data (FPLS; Wang & Santure, 2009). We analysed each simulated data set with all three methods, using no informative prior on sibship size. We also set allele frequencies as known, run length to medium and likelihood accuracy to "high". For all analyses, we included information about the identity of the mother of each half-sibling array. We note that inference based on pairwise relationships only is expected to perform more poorly than FL and FPLS, but we include it here for reference.

2.7 | *Antirrhinum majus* data

We applied the method to a sample of open-pollinated seeds collected in a hybrid-zone population of the snapdragon *A. majus* polymorphic for magenta and yellow pigmentation (Whibley et al., 2006). *Antirrhinum majus* has a closed mouth-like floral structure and is pollinated primarily by large bees which are strong enough to pull the flower open (Vargas, Ornosa, Ortiz-Sanchez, & Arroyo, 2010). Sampling of the progeny, mothers and candidate fathers and details of the genotyping procedure are described in detail by Ellis (2016). Briefly, we collected seed capsules from 96 mothers in July 2012, each containing many hundreds of seeds. We grew and collected tissue from a total of 1,468 seedlings from 57 of these families, with between 3 and 35 seedlings per family. We also sampled 2,128 adult plants, including maternal plants. DNA was extracted from leaf material from seedlings, maternal and paternal parents and subsequently genotyped at 120 SNPs by LGC Genomics. This SNP panel represents the 42 SNPs described by Ellis (2016) plus a further 27 SNPs designed using the same procedure. Parentage SNPs were chosen that showed as little spatial variation as possible, had minor allele frequencies close to 0.5 and that were at least 2 cM apart.

To estimate per-locus error rates associated with KASPR sequencing we repeated DNA extraction and genotyping for two independent tissue samples from each of 194 random adult plants. This generated 23,720 per-locus diploid genotypes, of which 0.13% differed between samples from the same individual. Based on this, we used genotyping error rates of 0.0013 for analyses with FAPS.

Unfortunately, the silica gel used to dry the offspring tissue did not have sufficient desiccating power, and the quality offspring DNA was highly variable. We have found that individuals with many loci that failed to amplify also had high genotype error rates at the remaining loci (David Luke Field, unpublished data). Rather than risk biasing statistical and biological conclusions through such errors, we applied a stringent data cleaning protocol prior to sibship analysis. We excluded 54 SNPs with more than 5% missing data and four with heterozygosity <0.2 or >0.75. We also excluded 736 offspring and 66 adult individuals with greater than 5% missing genotype data. After these data cleaning steps, the remaining 64 SNPs had on average 1.7% missing data in the offspring and 0.8% in the adults, and heterozygosity between 0.20 and 0.55.

We analysed these data in two stages. We first examined the largest family in the data set (20 offspring from mother L1872) to assess three factors which might indicate errors in genealogical inference. If assignment is accurate, we expect that the most probable candidates should be no more related to one another, nor show an increase in the proportion of missing genotypes than would be expected by a random draw from the population. Similarly, we would expect that the most probable pollen donors are found in close to the maternal plant, but that less likely candidates are drawn from the spatial distribution of candidates at random. We used FAPS to cluster family L1872 and identify most probable fathers for each offspring, accounting for uncertainty in sibship structure. We then compared distributions of pairwise relatedness, proportion of missing data and geographic location for the probable candidate fathers to those for the base population.

In a second analysis, we used a broader sample of families to infer the number of pollen donors that contribute to a seed capsule. To allow for direct comparison between families, we used only those 18 families with 17 or more offspring. We sampled 17 offspring at random from each family, leaving a total of 306 offspring. For each family, we estimated the posterior distributions of the number of contributing sires and the size of each full-sibship. We calculated 95% credible intervals based on the 2.5% and 9.7% percentiles of these distributions.

3 | RESULTS

3.1 | Simulations

3.1.1 | Method robustness

For the even-sibship-size mating scenario, all metrics of performance increased with the number of genotyped loci and decreased with increasing genotyping error rates and the number of candidate fathers in the sample (Figures 2 and 3). Using $\mu = 0.0015$ and 2,000 candidate males, 40 loci are sufficient to recover the true partition with >99% reliability, even for the largest samples of candidate males (Figure 2a,b). In 99.9% of cases where the true partition was not recovered, this was because FAPS identified an alternative partition structure with a higher likelihood of having generated the data than the true partition.

We found that 50 loci sufficed to identify full-sibling relationships and true sires with >95% posterior probability (Figure 2c). Under all simulation parameter sets, the distribution of family sizes remains centred on the true value, followed by a peak at family size of one, then one minus the true family size (Figure 3). Thus, when true full-siblings were assigned as half-siblings, this was in most cases due to a single individual being assigned to a singleton family.

Fractional analysis of paternity and sibships correctly inferred >99.9% of true half-sibling relationship across all parameter sets, even for cases with the fewest loci, highest errors and many candidate fathers. Partition structures that join two true full-sibships or otherwise overestimated family size had posterior probabilities very close to zero (Figure 3).

Simulation results were very similar in three other mating scenarios simulated (Figures S1–S4). One notable departure from the patterns described above is that for the single-family and all-half-siblings scenarios FAPS recovered the true partition in all simulated data sets (Figures S1 and S2). These scenarios represent the tip and base of the dendrogram (Figure 1). Furthermore, the probability p_{absent} that the true was not sampled was less than .001 under all mating scenarios.

3.1.2 | Reduced variance in larger families

In simulations investigating the effect of sibship size, true full-sibling relationships were recovered with >.98 posterior probability, regardless of family size (Figure 4). However, the variance in accuracy between samples decreased as family size increases, reflecting the increased information about sibling relationships with larger families.

3.1.3 | Little dependence on Monte Carlo draws

In all simulations, the proportion of probability space explored by the Monte Carlo sampling algorithm decreased as the number of candidate fathers increased (Figure 5, left-hand side). Increasing the number of Monte Carlo draws tended to increase the proportion of space explored. This effect was stronger when there were more candidate fathers in the sample. The increase was especially strong in the “many-full-sibships” and “reproductive-skew” scenarios, where there were many singleton offspring who could be compatible with multiple candidates (Figure 5c,g).

The accuracy of full-sibship inference also decreased with increased number of candidate fathers. However, there was no change in either p_{full} or p_{half} with increasing numbers of Monte Carlo draws.

3.1.4 | Comparison with Colony

All three analysis methods in Colony showed very high accuracy of full-sibling inference, even where the number of loci was very small (Figure 6a). FAPS showed substantially lower p_{full} than any Colony method for data sets with 30 or fewer loci, but was as accurate as Colony for data sets with 40 or more loci.

Fractional analysis of paternity and sibships, FL and FPLS also showed near total accuracy in inferring half-sibling relationships regardless of the number of loci, but the pairwise method showed very low p_{half} (Figure 6a). Further examination revealed that this was because the PLS method tended to group multiple true full-sibships into erroneous larger full-sibships.

The FL, PLS and FPLS completed analyses for all 300 data sets in 671.40, 322.43 and 465.47 min, respectively. FAPS required 2.05 min to create $\boldsymbol{G}$ matrices and perform hierarchical clustering for the same data sets on the same machine.

3.2 | Extensive polyandry in *A. majus*

For all of the 20 offspring in family L1872, there was a single candidate with a posterior probability of paternity >.96. These candidates were from a sample of seven independent most probable pollen donors. Given strong support for these candidates and for ease of presentation, we focus discussion on these individuals.

Most probable candidates were located between 32 and 79 m from the maternal plant (mean = 48 m). In contrast, second-most probable candidates were located between 102 and 1,367 m from the maternal plant (mean = 833 m; Figure S4). We found no evidence that most probable candidates were more related to one another than would be expected from a random draw from the pool of candidates (Figure S5), nor that most likely candidates had an unusually high proportion of missing SNP data (Figure S6).

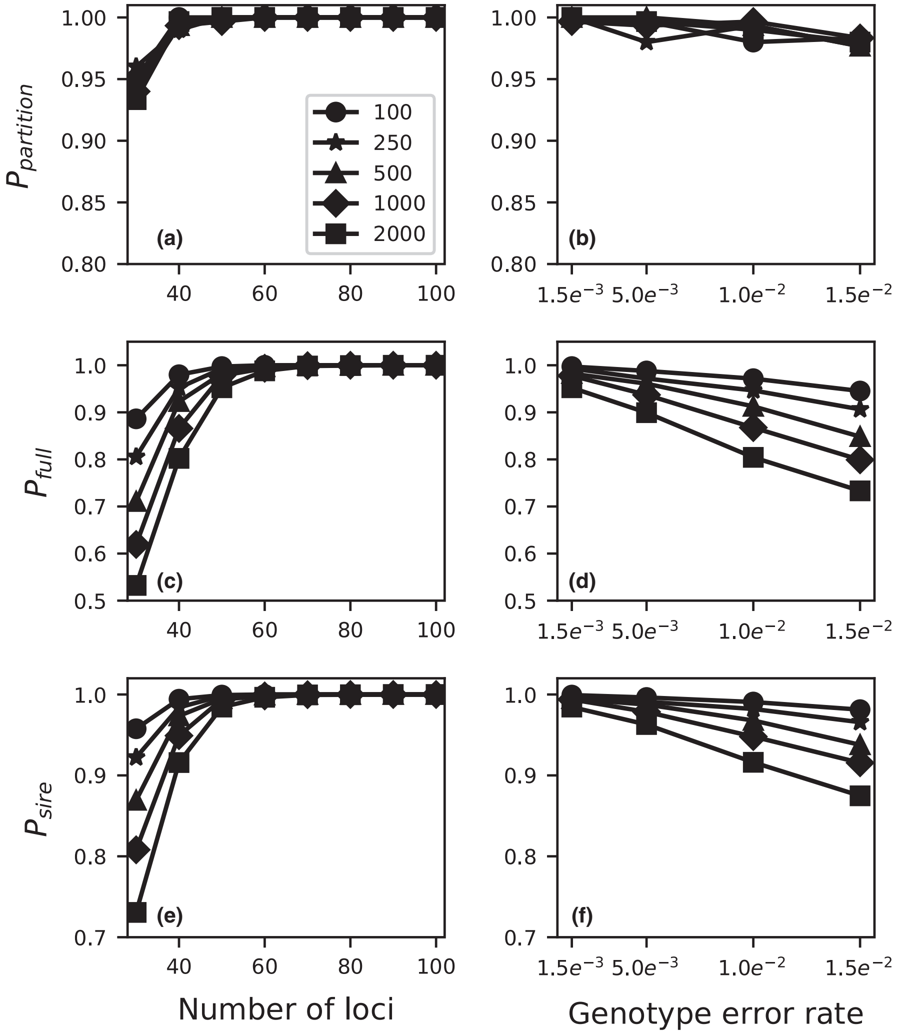

FIGURE 2 Fractional analysis of paternity and sibships performance for four families of five offspring. Subfigures show the probability of recovering the true partition (a, b), accuracy of full-sibling relationship reconstruction (c, d) and posterior probability of the each true sire on its offspring (e, f) as the number of typed loci (a, c, e), genotype error rate (b, d, e) and number of candidate fathers (see legend) vary

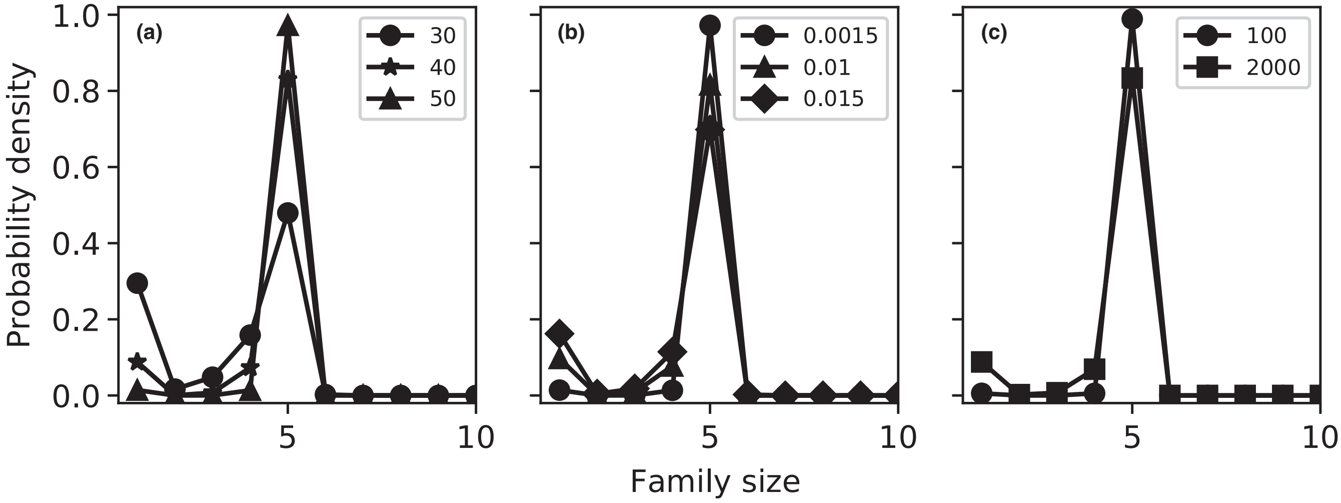

FIGURE 3 Distribution of inferred family sizes under the even-sibship-size mating scenario as number of loci (a), genotype error rate (b) and the number of candidate fathers (c) vary. For clarity, only an illustrative subset of the parameter sets are shown. In plots in which a single variable does not vary, plots show the cases for 50 loci, μ = 0.0015 and 250 candidates. The true family size is five in all cases

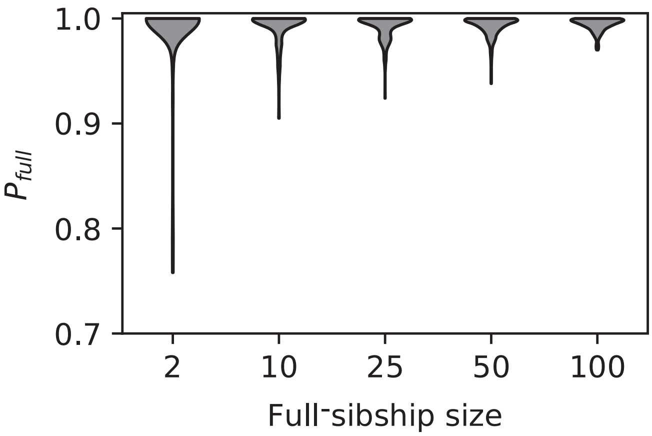

FIGURE 4 Viola plots showing the accuracy of pairwise full-sibling-relationship reconstruction as full-sibship size increases

We used FAPS to cluster 18 samples from wild-pollinated maternal families of *A. majus* into full-sibships. For all 306 offspring, the posterior probability that the true sire was unsampled was less than .001. The posterior distribution of family number implies that a sample of 17 offspring comprised between four and 15 full-sibling families of up to six offspring (Figure 7).

4 | DISCUSSION

In this study, we present a method for inferring sibship and paternity relationships from half-sibling arrays. We use hierarchical clustering to identify plausible ways to partition offspring into full-sibships and assess the support for different partition structures using Monte Carlo simulation. Given modest number of loci and realistic error rates, the method is as accurate as the algorithms implemented in Colony, but is faster by two orders of magnitude. The Python package FAPS also includes tools to extract biologically relevant information regarding sibship and paternity structure that automatically accounts for uncertain about the exact pedigree. As such FAPS represents an accurate and efficient tool that allows for robust and efficient biological inference from genealogies.

Using realistic sample sizes, a modest number of SNP loci error rates, FAPS performs well under a variety of contrasting sibship structures. This is true even when samples of offspring and candidate fathers are large, which previous work has found can be problematic for sibship assignment (Almudevar & Anderson, 2012). In particular, FAPS never falsely assigned two full-siblings to be half-siblings. When there were errors in sibship assignment, single offspring tended to splinter into families on their own. Most such errors were due to individuals being assigned to a family on their own, rather with other individuals in a larger family (Figure 3). These errors typically occurred when an unrelated candidate male had a higher likelihood of paternity than did the true sire due to stochasticity in Mendelian sampling. This phenomenon has been previously noted for both paternity (Thompson, 1976) and sibship (Butler, Field, Herbinger, & Smith, 2004) assignment problems. In the overwhelming majority of cases where FAPS erred, this was due to the existence of an alternative partition with higher likelihood than the true partition, rather than a failure of the clustering algorithm to detect the true partition. As such, the incorrect partitions identified by FAPS is in fact more consistent with the data than the true genealogy.

We found Monte Carlo simulation to be an efficient approach to estimating the likelihood of a partition, while excluding the possibility that multiple full-sibships share a father. A drawback of this scheme is that there is no way to know how much of the probability space has been sampled for a given number of Monte Carlo draws. Increasing the number of Monte Carlo draws increased the proportion of probability space that could be explored, especially in samples with many candidate males, or smaller full-sibship sizes (Figure 5). This is not surprising, given that, in these scenarios, the space of possible configurations is much larger, and there is less information among full-siblings to rule out unrelated candidate fathers. However, the number of draws had almost no effect on the accuracy of sibship inference (Figure 5, right-hand side). This indicates that the most likely configurations can be sampled with a small number of draws and that the accuracy of FAPS does not depend on the parameter choice for the number of Monte Carlo draws.

Our hierarchical clustering algorithm relies on a matrix of probabilities that pairs of offspring are full-siblings. As well as having low statistical power (Wang, 2007), pairwise sibship measures can be problematic in that any pair of three individuals can be compatible as full-siblings sired by separate fathers, but incompatible as a single family sired by a single father. Several lines of evidence indicate that this issue is not a significant concern for this method. Firstly, the Monte Carlo sampling scheme explicitly evaluates the likelihood that the whole sibship was sired by individual candidate fathers. As noted above, simulations demonstrate that this returns accurate sibship and paternity configurations (Figures 2–4 and S1–S4). Finally, FAPS dramatically outperforms the pairwise-likelihood method implemented in Colony (Figure 6). These observations indicate that any inherent weakness in using the heuristic pairwise metric has negligible negative impact on the accuracy of this method.

Our analysis of open-pollinated *A. majus* seed capsules found evidence for a large number of small full-sibling families in a half-sibling array. As a single seed capsule can contain hundreds of seeds, our samples of 17 offspring per array represent only a tiny fraction of the total seeds in the array. It is therefore likely that these full-sibships are substantially larger than implied in Figure 7. Nevertheless, this demonstrates that a single seed capsule contains offspring from a large number of pollen donors, either through many independent pollinator visits or through fewer visits by pollinators delivering mixed pollen loads. As pollinator behaviour is a crucial mediator of gene flow among flowering plants, we are using these paternity data in ongoing work to investigate the interaction between flower colour and pollinator behaviour in the maintenance of this hybrid zone.

Closer examination of a single large family found that the 20 offspring in family L1872 could be assigned to seven pollen donors with high posterior probability. These donors were no more related

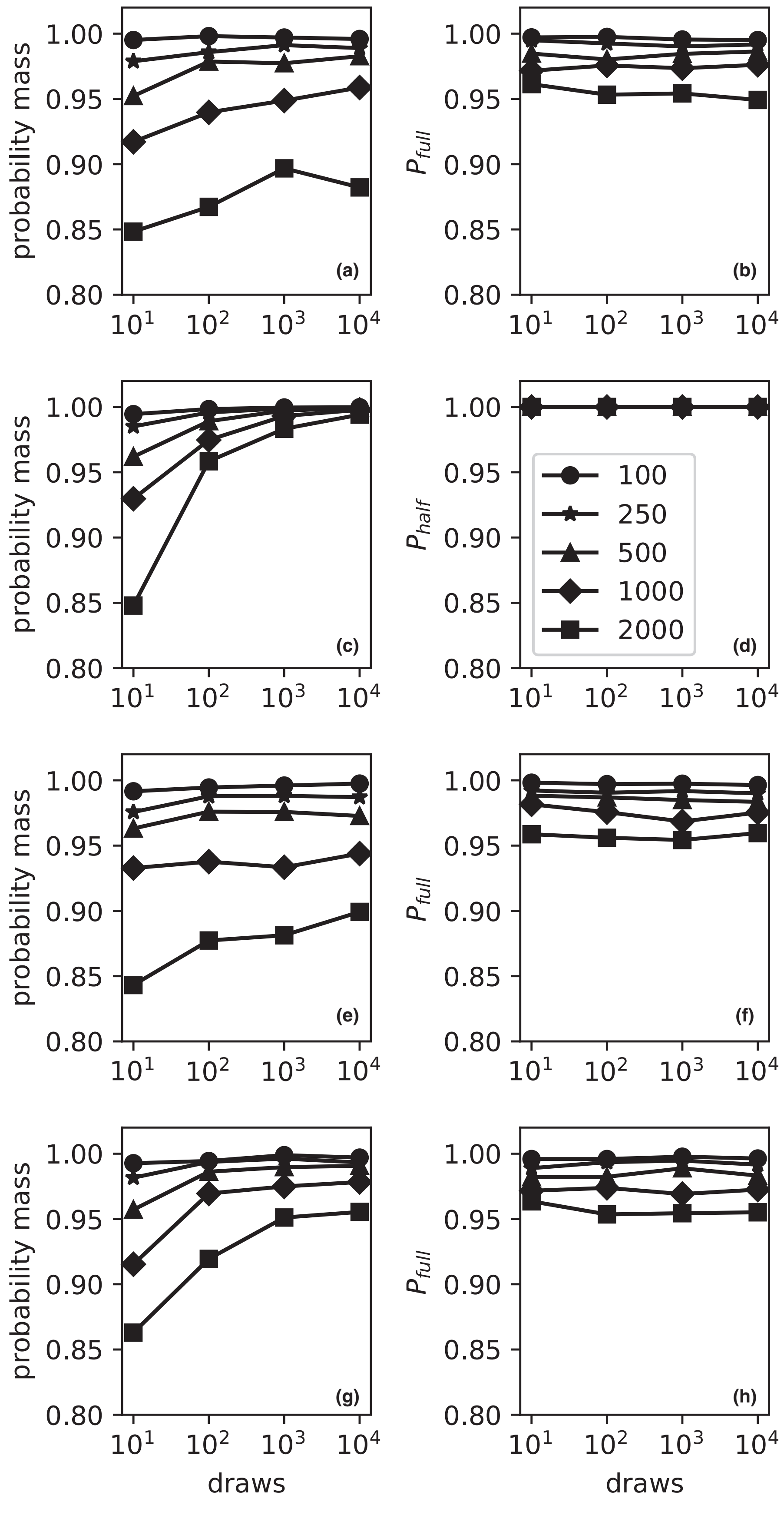

FIGURE 5 Effect of the number of Monte Carlo draws on the proportion of probability space sampled (a, c, e, g) and accuracy of sibship reconstruction (b, d, f, g). Legend indicates number of candidate fathers. Subplots show scenarios for even sibship sizes (a–b), a single family (c–d) all half-siblings (e–f) reproductive skew (g–h)

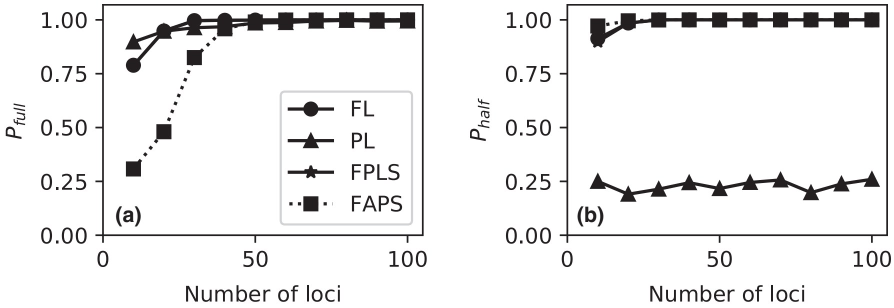

FIGURE 6 Sibship inference using Colony and Fractional analysis of paternity and sibships (FAPS). Curves show reconstruction of full-sibling (a) and half-sibling (b) relationships using the three methods in Colony (FL, PLS, FPLS) and FAPS

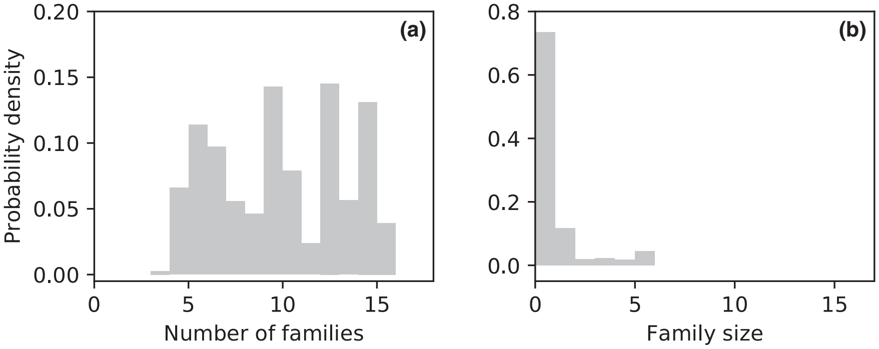

FIGURE 7 Sibship and inference in 18 *Antirrhinum majus* half-sibling arrays showing posterior distributions of the number of full-sibships (a) and the size of each full-sibship (b). Shaded regions show 95% credible intervals

to one another nor have substantially poorer quality genotype data than would be expected by chance. Furthermore, they were within close proximity to the maternal plant, as one would expect if pollen is transported by foraging bees (Ashley, 2010). In contrast, the second-most probable set of candidate fathers was located across the population, as we would expect for unrelated individuals drawn at random from the pool of candidates. These observations indicate that the most probable fathers are the true sires and that results from FAPS for the broader survey of 18 families are robust.

Inference of the relationships between individuals in natural populations is an important technique for understanding patterns of gene dispersal and selection in the wild (Ashley, 2010; Pemberton, 2008). The method presented here represent a useful tool to infer mating patterns from half-sibling arrays that accounts for uncertainty about exact relationships. Because the method requires only a likelihood of paternity for mother–offspring–father triplets, it does not depend on marker type or genetic system, provided this likelihood can be calculated. We have focused on the case where one mother is known and individuals are genotyped using biallelic SNPs. Nevertheless, it could in principle be easily extended to the assignment of parent pairs simply by substituting an appropriate likelihood function to calculate **G** (Meagher & Thompson, 1986), although the greater number of possible parents would require more markers and more intense computation. Moreover, it is equally applicable to any marker type (Anderson & Garza, 2006; Jones & Ardren, 2003) or organisms with polyploid inheritance (Wang & Scribner, 2014), provided that it is possible to estimate likelihoods of paternity.

ACKNOWLEDGEMENTS

We thank a large number of field volunteers involved in data collection of the *A. majus* population and Matthew Couchman and Harald Ringbauer for their help in processing and cleaning these data. We are also grateful to four anonymous reviewers whose comments significantly improved the quality of the manuscript. This work was partially supported by ERC grant 250152 to NHB.

DATA AVAILABILITY

The FAPS package and documentation are available at https://github.com/ellisztamas/faps and from the PyPi package repository. Genotype and GPS data for the *A. majus* data set, as well as scripts to run the simulations, are available from the IST Austria data repository at https://doi.org/10.15479/at:ista:95.

AUTHOR CONTRIBUTIONS

The method was conceived and simulations and analysis of family structure in *A. majus* were performed by T.J.E. Sampling was arranged, processing of the *A. majus* population was carried out and the SNP panel was designed by D.L.F. The manuscript was written by T.J.E., D.L.F. and N.H.B.

ORCID

Thomas James Ellis http://orcid.org/0000-0002-8511-0254

REFERENCES

Almudevar, A., & Anderson, E. C. (2012). A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. *Molecular Ecology Resources*, *12*(1), 164–178. https://doi.org/10.1111/j.1755-0998.2011.03061.x

Anderson, E. C., & Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, *172*(4), 2567–2582.

Anderson, E. C., & Ng, T. C. (2016). Bayesian pedigree inference with small numbers of single nucleotide polymorphisms via a factor-graph representation. *Theoretical Population Biology*, *107*, 39–51. https://doi.org/10.1016/j.tpb.2015.09.005

Ashley, M. V. (2010). Plant parentage, pollination, and dispersal: How DNA microsatellites have altered the landscape. *Critical Reviews in Plant Sciences*, *29*(3), 148–161. https://doi.org/10.1080/07352689.2010.481167

Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, *18*(10), 503–511. https://doi.org/10.1016/S0169-5347(03)00225-8

Butler, K., Field, C., Herbinger, C., & Smith, B. (2004). Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Molecular Ecology*, *13*(6), 1589–1600. https://doi.org/10.1111/j.1365-294X.2004.02152.x

Devlin, B., Roeder, K., & Ellstrand, N. (1988). Fractional paternity assignment: Theoretical development and comparison to other methods. *Theoretical and Applied Genetics*, *76*(3), 369–380.

Ellis, T. J. (2016). *The role of pollinator-mediated selection in the maintenance of a flower colour polymorphism in an Antirrhinum majus hybrid zone*. PhD thesis, Institute of Science and Technology Austria, Klosterneuburg.

Emery, A., Wilson, I., Craig, S., Boyle, P., & Noble, L. (2001). Assignment of paternity groups without access to parental genotypes: Multiple mating and developmental plasticity in squid. *Molecular Ecology*, *10*(5), 1265–1278. https://doi.org/10.1046/j.1365-294X.2001.01258.x

Hadfield, J., Richardson, D., & Burke, T. (2006). Towards unbiased parentage assignment: Combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology*, *15*(12), 3715–3730. https://doi.org/10.1111/j.1365-294X.2006.03050.x

Jones, A. G., & Ardren, W. R. (2003). Methods of parentage analysis in natural populations. *Molecular Ecology*, *12*(10), 2511–2523. https://doi.org/10.1046/j.1365-294X.2003.01928.x

Jones, B., Grossman, G. D., Walsh, D. C., Porter, B. A., Avise, J. C., & Fiumera, A. C. (2007). Estimating differential reproductive success from nests of related individuals, with application to a study of the mottled sculpin, cottus bairdi. *Genetics*, *176*(4), 2427–2439. https://doi.org/10.1534/genetics.106.067066

Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, *10*(1), 6–30. https://doi.org/10.1111/j.1755-0998.2009.02778.x

Marshall, T., Slate, J., Kruuk, L., & Pemberton, J. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, *7*(5), 639–655. https://doi.org/10.1046/j.1365-294x.1998.00374.x

McKinney, W. (2010). Data structures for statistical computing in python. In van der Walt S. & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56).

Meagher, T. R. (1986). Analysis of paternity within a natural population of *Chamaelirium luteum*. 1. Identification of most-likely male parents. *American Naturalist*, *128*(2), 199–215. https://doi.org/10.1086/284554

Meagher, T. R., & Thompson, E. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology*, *29*(1), 87–106. https://doi.org/10.1016/0040-5809(86)90006-7

Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, *53*(9), 1–18.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86–97.

Neff, B. D., Repka, J., & Gross, M. R. (2001). A bayesian framework for parentage analysis: The value of genetic and other biological data. *Theoretical Population Biology*, *59*(4), 315–331. https://doi.org/10.1006/tpbi.2001.1520

Nielsen, R., Mattila, D. K., Clapham, P. J., & Palsbøll, P. J. (2001). Statistical approaches to paternity analysis in natural populations and applications to the north atlantic humpback whale. *Genetics*, *157*(4), 1673–1682.

Pemberton, J. (2008). Wild pedigrees: The way forward. *Proceedings of the Royal Society of London B*, *275*(1635), 613–621. https://doi.org/10.1098/rspb.2007.1531

Roeder, K., Devlin, B., & Lindsay, B. G. (1989). Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, *45*(2), 363–379. https://doi.org/10.2307/2531483

Sieberts, S. K., Wijsman, E. M., & Thompson, E. A. (2002). Relationship inference from trios of individuals, in the presence of typing error. *The American Journal of Human Genetics*, *70*(1), 170–180. https://doi.org/10.1086/338444

Sokal, R. R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

Thomas, S. C., & Hill, W. G. (2002). Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetical Research*, *79*(03), 227–234.

Thompson, E. (1976). A paradox of genealogical inference. *Advances in Applied Probability*, *8*(4), 648–650. https://doi.org/10.2307/1425927

Thompson, E., & Meagher, T. (1987). Parental and sib likelihoods in genealogy reconstruction. *Biometrics*, *43*, 585–600. https://doi.org/10.2307/2531997

Vargas, P., Ornosa, C., Ortiz-Sanchez, F., & Arroyo, J. (2010). Is the occluded corolla of *Antirrhinum* bee-specialized? *Journal of Natural History*, *44*(23–24), 1427–1443. https://doi.org/10.1080/00222930903383552

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics*, *166*(4), 1963–1979. https://doi.org/10.1534/genetics.166.4.1963

Wang, J. (2007). Parentage and sibship exclusions: Higher statistical power with more family members. *Heredity*, *99*(2), 205–217. https://doi.org/10.1038/sj.hdy.6800984

Wang, J. (2012). Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics*, *191*(1), 183–194. https://doi.org/10.1534/genetics.111.138149

Wang, J., & Santure, A. W. (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, *181*(4), 1579–1594. https://doi.org/10.1534/genetics.108.100214

Wang, J., & Scribner, K. T. (2014). Parentage and sibship inference from markers in polyploids. *Molecular Ecology Resources*, *14*(3), 541–553. https://doi.org/10.1111/1755-0998.12210

Whibley, A. C., Langlade, N. B., Andalo, C., Hanna, A. I., Bangham, A., Thébaud, C., & Coen, E. (2006). Evolutionary paths underlying flower color variation in *Antirrhinum*. *Science*, *313*(5789), 963–966. https://doi.org/10.1126/science.1129161

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Ellis TJ, Field DL, Barton NH. Efficient inference of paternity and sibship inference given known maternity via hierarchical clustering. *Mol Ecol Resour*. 2018;00:1–12. https://doi.org/10.1111/1755-0998.12782

APPENDIX 1

Likelihood of paternity

The likelihood that a male is the true father of an offspring is given by the probability λ_l of observing the offspring genotype given the maternal and paternal alleles at locus *l*, multiplied across each locus, such that $L(o_i|f_j, m) = \prod_l \lambda_l$ (Devlin et al., 1988; Meagher, 1986; Meagher & Thompson, 1986). This formulation assumes that all loci are unlinked, which may not hold for SNP data. Anderson and Garza (2006) found that linkage causes the SNP panel to behave as if there were fewer loci than had been typed, but that this effect was fairly minor.

One source of genotyping error are point mutations, where a haploid genotype is observed to be allele A when it is actually allele B. We follow Anderson and Garza (2006) in summing over all possible maternal, paternal and offspring alleles, weighted by the probability that each is the true genotype given error rate ϵ_1. This rate can be estimated through repeat genotyping of the same individuals.

A further source of error occurs where one or more loci fail to amplify for an individual, leaving missing data for that locus. Failing to account for these missing data causes candidate males with many failed loci to have high likelihoods of paternity, because the calculation of $\Pr(o_i|f_j, m)$ multiplies over fewer loci. We account for these errors by correcting for the *v* loci which amplified successfully for the maternal, paternal and offspring genotypes, giving $\Pr(o_i|f_j, m) = \prod_l \lambda_l^{1/v}$, or equivalently $\log g_{ik} \propto 1/v \sum_l \log \lambda_l$. This dropout rate, ϵ_2, can be observed directly from the data.

Incomplete sampling of males

In real data sets, it is unlikely that every male can be sampled, and therefore, some offspring may have paternal genotypes not found in *F*. Following Nielsen et al. (2001), we modify **G** to account for the probability θ_i that the father of the *i*th offspring has been sampled and probability $\Pr(o_i|m, \mathbf{a})$ that an individual is the offspring of an unsampled father with alleles drawn at random from the vector of local allele frequencies **a**. Thus we have

$$g_{ij} = \frac{\theta_i \Pr(o_i|f_j, m)}{\sum_j \theta_i \Pr(o_i|f_j, m) + (1-\theta_i)\Pr(o_i|\mathbf{a})}.$$

To ensure rows in **G** sum to one, we also append each row in **G** with

$$g_{ia} = \frac{(1-\theta_i)\Pr(o_i|\mathbf{a})}{\sum_k \theta_i \Pr(o_i|f_k) + (1-\theta_i)\Pr(o_i|\mathbf{a})}.$$

Individuals with large terms for g_{ia} may either be full-siblings sharing a single unsampled father or else half-siblings with different unsampled fathers. It is difficult to distinguish these hypotheses in the absence of further information. When estimating $\Pr(T_c|\mathbf{G})$, we therefore keep the requirement that no two sibships can share a father in *F*, but do allow two sibships to share an unsampled father.

It is assumed that the sample of candidate males is sufficiently large to give a representative estimate of **a** or that estimates are available from other sources. If it is expected that only a small proportion of the candidate males has been sampled, it would be appropriate to consider inference of sibship relationships without parental information or accurate allele frequency data, such as Colony (Wang, 2004).

By including g_{ia} in **G**, it is straightforward to account for unsampled fathers in biological inference. A missing father is treated like any other father, and the probability of a missing father is automatically incorporated into **G**. This will not bias inference about the relationship between fecundity and phenotypes provided at the phenotype of the missing fathers are random.