Title

Joint Estimation of Pedigrees and Effective Population Size Using Markov Chain Monte Carlo.

Journal

Genetics, 212(3)

ISSN

0016-6731

Authors

Ko, Amy
Nielsen, Rasmus

Publication Date

2019-07-01

DOI

10.1534/genetics.119.302280

Joint Estimation of Pedigrees and Effective Population Size Using Markov Chain Monte Carlo

Amy Ko*,[1] and Rasmus Nielsen*,†,‡

*Department of Integrative Biology and †Department of Statistics, University of California, Berkeley, 94720California, and ‡Museum of Natural History, University of Copenhagen, 1123 Denmark

ORCID IDs: 0000-0001-7993-5463 (A.K.); 0000-0003-0513-6591 (R.N.)

ABSTRACT Pedigrees provide the genealogical relationships among individuals at a fine resolution and serve an important function in many areas of genetic studies. One such use of pedigree information is in the estimation of the short-term effective population size (N_e), which is of great relevance in fields such as conservation genetics. Despite the usefulness of pedigrees, however, they are often an unknown parameter and must be inferred from genetic data. In this study, we present a Bayesian method to jointly estimate pedigrees and N_e from genetic markers using Markov Chain Monte Carlo. Our method supports analysis of a large number of markers and individuals within a single generation with the use of a composite likelihood, which significantly increases computational efficiency. We show, on simulated data, that our method is able to jointly estimate relationships up to first cousins and N_e with high accuracy. We also apply the method on a real dataset of house sparrows to reconstruct their previously unreported pedigree.

KEYWORDS pedigree inference; effective population size; Markov Chain Monte Carlo

PEDIGREES are fundamental in many areas of genetic studies. Pedigree structure can be used to study the social organization of a population, such as the degree of polygamy and the offspring distribution among mothers and fathers (Blouin 2003). In conservation genetics, pedigrees provide a way to design an appropriate breeding scheme by preventing inbreeding between close relatives. Other uses of pedigree information include estimating heritability of quantitative traits (Vinkhuyzen *et al.* 2013), controlling for cryptic relatedness in association studies (Voight and Pritchard 2005; Eu-ahsunthornwattana *et al.* 2014), and pedigree-based association studies (Ott *et al.* 2011). Furthermore, the genealogical history embedded in pedigrees can be used to estimate demographic parameters for the recent past, such as the short-term effective population size (N_e) (Wang 2009). However, most population genetic models are based on Kingman's coalescent (Kingman 1982a,b,c), which is a poor approximation of the genealogical process for time frames shorter than $log_2 N$, where N is the population size (Wakeley *et al.* 2012, 2016). Pedigrees, which provide a finer resolution on the genealogical history of the samples than the coalescent, may therefore be more appropriate to use for estimating demographic parameters of the very recent past.

Despite the importance of pedigrees in genetic analyses, they are often a missing parameter. To address this problem, many methods have been developed to estimate pedigrees from genetic data. Existing methods fall broadly into two categories: those that estimate pairwise relationships only (Thompson 1975; McPeek and Sun 2000; Smith *et al.* 2001; Sun *et al.* 2001; Milligan 2003; Sun and Dimitromanolakis 2014) and those that aim to reconstruct the entire pedigree (Thomas and Hill 2000; Almudevar 2003; Wang 2004, 2012; Hadfield *et al.* 2006; Gasbarra *et al.* 2007; Cowell 2009, 2013; Riester *et al.* 2009; Wang and Santure 2009; Kirkpatrick *et al.* 2011; Almudevar and Anderson 2012; Cussens *et al.* 2013; He *et al.* 2013; Staples *et al.* 2014, 2016; Anderson and Ng 2016; Ko and Nielsen 2017; Ramstetter *et al.* 2018). Although pairwise methods are computationally fast, estimated pairwise relationships do not necessarily translate to the correct pedigree, as piecing together pairwise relationships may not produce a valid

doi: https://doi.org/10.1534/genetics.119.302280
Manuscript received December 6, 2018; accepted for publication May 16, 2019; published Early Online May 22, 2019.
Supplemental material available at FigShare: https://doi.org/10.25386/genetics.8079953.
[1]Corresponding author: Department of Integrative Biology, University of California, Berkeley, Valley Life Science Bldg., Harmon Way, Berkeley, CA 94720. E-mail: amyko@berkeley.edu

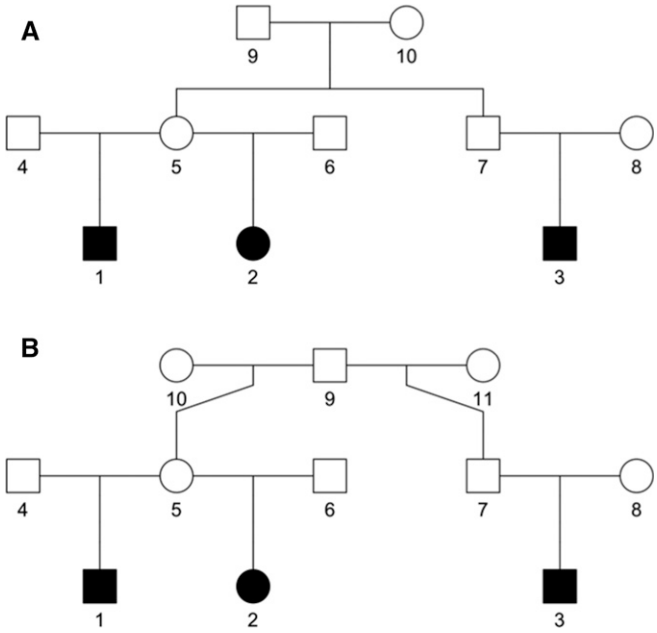

Figure 1 An example output pedigrees for three sampled individuals (shaded) from a dataset in Simulation A. Sex of the unsampled individuals (unshaded) are unknown but are drawn in for illustration only. (A) Pedigree with the highest estimated posterior probability ($P = 0.55$). (B) Pedigree with the second highest estimated posterior probability ($P = 0.45$). The true pedigree is shown in (A).

Table 1 Pairwise prediction accuracy for simulation A (SNPs), aggregated over 50 independent datasets

		Predicted				
		FS	HS	UR	FC	HC
True	FS	106	0	0	0	0
	HS	0	136	0	1	0
	UR	0	0	59,996	0	4
	FC	1	0	0	445	32
	HC	0	0	0	4	526

Two-generation inference by MCMC.

pedigree. Furthermore, because the coefficient of variation in genome sharing between two individuals becomes larger as the relationship becomes more distant (Hill and Weir 2011), distinguishing competing relationships from each other becomes increasingly difficult. Methods that estimate the entire pedigree have an advantage in this regard. Several studies have shown that the accuracy of pairwise relationship inference can be improved by considering all relationships in the sample simultaneously and resolving uncertain relationships in the context of other individuals (Staples *et al.* 2014; Ko and Nielsen 2017; Ramstetter *et al.* 2018). Furthermore, the estimated pedigree is valid by construction (see *Appendix* for the conditions for a valid pedigree), which can then be used to study population parameters of interest, such as the variance in offspring distribution.

Existing pedigree reconstruction methods, however, are limited in their scope due to the inherent difficulty in pedigree inference. First, the likelihood computation of a pedigree is expensive, as it scales exponentially either in the number of individuals in the pedigree (Lander and Green 1987) or in the number of markers analyzed (Elston and Stewart 1971). Second, the number of possible pedigrees for a given number of individuals is enormous, much greater than the number of phylogenetic trees (Steel and Hein 2006; Thatte and Steel 2008), which makes exploring the pedigree space computationally challenging, even for a small number of individuals.

In this study, we present a pedigree inference method that addresses the difficulties of pedigree inference. First, we use the composite likelihood developed in Ko and Nielsen (2017) to make the likelihood computation efficient for a large number of markers and individuals. Second, we use Markov Chain Monte Carlo (MCMC) (Hastings 1970) to sample pedigrees from high probability regions, circumventing the need to enumerate all possible pedigrees. Our method is different in several important ways from previous methods (see, *e.g.*, Wang 2012; Staples *et al.* 2014; Ko and Nielsen 2017) that also use composite likelihoods and sampling algorithms to explore the pedigree space. These previous methods take a maximum likelihood approach and produce a list of pedigrees with highest likelihoods, and do not provide a principled way to compute the uncertainty of the estimated pedigrees. In contrast, our method casts the problem in a Bayesian framework, and estimates the posterior probability distribution of the parameters, which, in turn, quantifies the uncertainty in parameter estimation.

Furthermore, by assigning a prior, which is a function of population parameters that govern the mating behavior of the population, to the pedigrees, we can estimate these parameters jointly with the pedigree. In particular, we focus on estimating the short-term N_e, a key parameter quantifying the level of genetic drift and inbreeding in the current population. Various approaches have been developed for estimating the short-term N_e, including methods based on relatedness, heterozygosity excess, linkage disequilibrium (LD), or changes in allele frequency over time (Wang *et al.* 2016). Our pedigree-based approach for estimating N_e is most closely related to the estimation method based on the frequency of siblings in a sample by Wang (2009), which was shown to be more accurate and robust than other approaches. A review by Hendricks *et al.* (2018) gives an overview of various methods for N_e estimation in the context of conservation genetics.

In our method, we assume that all sampled individuals belong to a single generation and infer pedigrees going up to two generations back in time (*i.e.*, up to first cousins). Furthermore, we assume that the population is outbred with nonoverlapping generations, and the pedigrees do not contain cycles other than those caused by full sibling relationships. We validate our method on simulated data and show that it can estimate relationships and N_e accurately. Furthermore, we apply our method on a real dataset containing a sample of house sparrows to reconstruct their previously unreported pedigree.

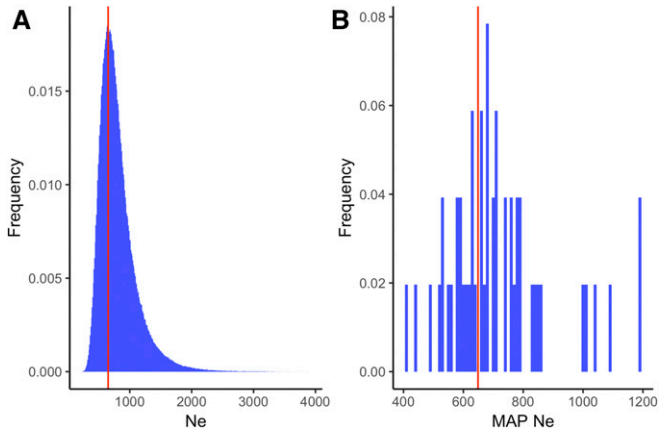

Figure 2 (A) Estimated posterior distribution of *Ne* from MCMC samples aggregated over 50 datasets in Simulation A. (B) Distribution of maximum a posterior (MAP) N_e for the 50 datasets in Simulation A. The red vertical line in each panel corresponds to the true value of the parameter.

Table 2 Pairwise prediction accuracy for simulation A (SNPs), aggregated over 50 independent datasets

		Predicted[a]		
		FS	HS	UR
True	FS	106	0	0
	HS	0	137	0
	UR	0	0	60,000
	FC	0	117	360
	HC	0	0	530

One-generation inference by MCMC.
[a] The likelihoods were computed without using the linkage information between markers to make the likelihood computation comparable to COLONY's.

Materials and Methods

Bayesian inference of pedigrees and mating parameters

Our method aims to estimate the joint posterior distribution of pedigrees and mating parameters. Let n be the sample size, H the pedigree of the sample, θ the set of mating parameters for the population, and $X = (X_1, \ldots, X_n)$ the set of genotype vectors for the n individuals. Then, the joint posterior probability of H and θ can be written as

$$Pr(H, \theta|X) \propto Pr(X|H)Pr(H|\theta)Pr(\theta), \quad (1)$$

where $Pr(X|H)$ is the likelihood of the pedigree, $Pr(H|\theta)$ is the prior for the pedigree under a mating model parameterized by θ, and $Pr(\theta)$ is the hyperprior on the mating parameters. We describe below in more detail how to compute each of these component terms.

Composite likelihood: As discussed in the Introduction, computing the likelihood of a pedigree, $Pr(X|H)$, is computationally prohibitive for even a moderately large set of markers or individuals. We therefore approximate the likelihood with the composite likelihood introduced in Ko and Nielsen (2017) to make the computation more efficient. The composite likelihood is based on the marginal pairwise likelihoods, which we describe briefly here.

The composite likelihood of a local pedigree H_l for k sampled individuals is given by

$$CL(H_l) = \begin{cases} P(X_i), & \text{if } k = 1 \\ \dfrac{\prod_{(i,j)\in H_l} P(X_i, X_j|R_{i,j})}{\prod_{i\in H_l} P(X_i)^{k-2}}, & \text{otherwise} \end{cases} \quad (2)$$

where $R_{i,j}$ is the relationship between individuals i and j induced by pedigree H_l. If the pedigree contains a single individual (*i.e.*, $k = 1$), then the composite likelihood is simply the probability of observing the individual's genotypes (*i.e.*, product of the genotype frequencies). For $k > 1$, the composite likelihood is the product of the pairwise likelihoods, scaled by the marginal likelihoods of the individuals. That is, since each individual appears $k - 1$ times in the numerator, we divide the numerator by the marginal likelihood of each individual $k - 2$ times. A previous study by Ko and Nielsen (2017) showed that the composite likelihood scales similarly to the full likelihood on simulated data and has sensible asymptotic properties, making it a good approximation for the full likelihood. The full composite likelihood for a set of local pedigrees, H, is the product of the composite likelihood for each local pedigree, H_l.

We precompute and store in memory the pairwise likelihoods $Pr(X_i, X_j|R_{i,j})$ for each pair (i,j) for a specified set of pairwise relationships. For pedigrees going up to two generations back in time, this set includes full siblings, half siblings, full first cousins, half first cousins, and unrelated. The pairwise likelihoods can be computed efficiently using the method described in Weir *et al.* (2006) for unlinked markers or by Albrechtsen *et al.* (2009) for linked markers. The pairwise likelihoods can then be accessed from memory to compute the composite likelihood efficiently.

Prior: For the prior on the pedigrees, $Pr(H|\theta)$, we used a modified version of the mating model introduced in Gasbarra *et al.* (2005). The model is defined by three parameters: α, β, and N, which we describe in more detail below. The modified version of the model does not change the original equations in Gasbarra *et al.* (2005), but affects the interpretations of the mating parameters, which will be discussed below.

The probability of a pedigree under this mating model is most naturally described by the procedure by which each child stochastically chooses its mother and father. We assume a homogeneous population of constant size N with nonoverlapping generations and equal proportions of males and females (*i.e.*, $N/2$ males and $N/2$ females). Let n be the number of children in the current generation. One by one, each child chooses a parental pair (f, m), where $f \in \{1, 2, \ldots, N/2\}$ and $m \in \{1, 2, \ldots, N/2\}$.

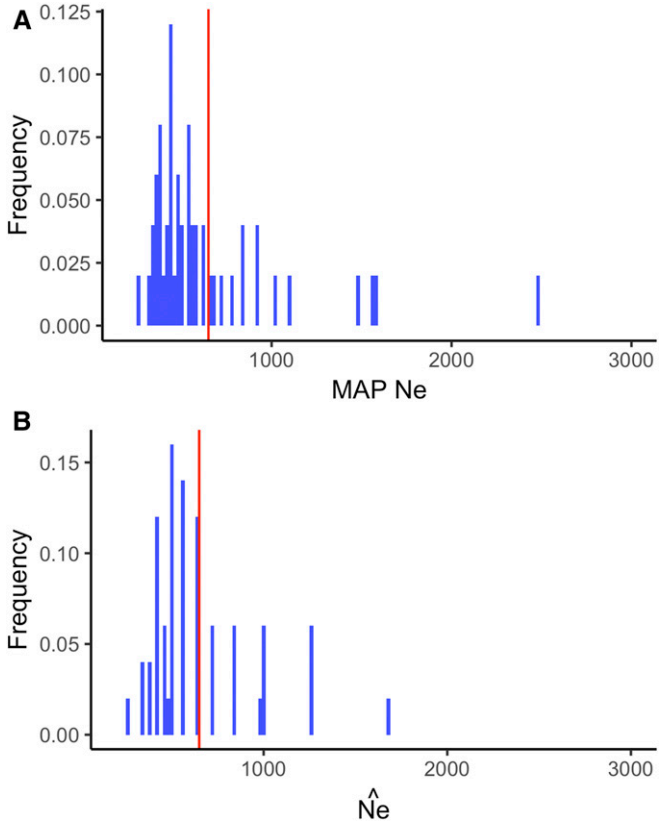

Figure 3 (A) Distribution of MAP N_e by MCMC, where the pedigree inference was restricted to one generation and the likelihood computation assumed independent markers. (B) Distribution of N_e estimates by COLONY based on full likelihood computation with independent markers and nonrandom mating.

Table 3 Pairwise prediction accuracy for simulation A (SNPs), aggregated over 50 independent datasets

		Predicted[a]		
		FS	HS	UR
True	FS	106	0	0
	HS	0	137	0
	UR	0	0	60,000
	FC	0	106	371
	HC	0	0	530

One-generation inference by COLONY.
[a]Inference was based on the full likelihood method under the assumption of independent markers.

Let $C_f(k)$ be the number of children that mother f has after the first k children have chosen their parents. Then the probability that the $(k+1)$th child chooses mother f is given by

$$\frac{\alpha + C_f(k)}{\alpha(N/2) + k}, \tag{3}$$

where α is a parameter that controls the offspring distribution among mothers in the population. A small value of α corresponds to the mating model where a few mothers have many offspring, whereas a large value of α corresponds to the model where children are distributed more evenly among all mothers.

After selecting mother f, the child chooses a father next. Let $C_{fm}(k)$ be the number of children that parental pair (f, m) has after the first k children have chosen their parents. Then the probability of the $(k+1)$th child choosing father m is given by

$$\frac{\beta + C_{fm}(k)}{\beta(N/2) + C_f(k)}, \tag{4}$$

where β is a parameter that governs the degree of polygamy of fathers. If β is small, then the child is more likely to choose father m if the father already shares other offspring with the child's mother, f (*i.e.*, parental pairs tend to stay monogamous). On the other hand, $\beta = \infty$ corresponds to the case where the child chooses a father at random (*i.e.*, random mating model).

After all n children in the current generation have chosen their parents, we continue recursively backward in time by treating the chosen mothers and fathers in the current stage as the offspring for the next stage. Using this sequential sampling scheme, we can compute $Pr(H|\theta)$, where $\theta = (\alpha, \beta, N)$.

Furthermore, we can relate the mating parameters α, β, and N to the effective population size, N_e, using the formula derived in Gasbarra *et al.* (2005).

For the hyperprior, $P(\theta)$, we assume a uniform distribution for each of the parameters in θ. For instance, we assume $\alpha \sim U(\alpha_{min}, \alpha_{max})$ for some fixed α_{min} and α_{max}. We treat β and N in a similar way.

Finally, we combine the composite likelihood, prior, and hyperprior to approximate the joint posterior distribution of H and θ with

$$CL(H)Pr(H|\theta)Pr(\theta) \tag{5}$$

MCMC: To explore the vast parameter space in a computationally feasible way, we use MCMC to sample from the posterior distribution of H and θ, approximated by Equation 5.

We represent the pedigree for a sample of individuals as an undirected graph, where a node corresponds to an individual with a particular sex (*i.e.*, male or female), and an edge represents a parent–offspring relationship. Individual i in the graph is not necessarily represented in the sample; but, if it is sampled, the node is associated with a genotype vector, X_i. A more detailed description of the graph representation of pedigrees and the conditions for a valid pedigree is provided in *Appendix*.

The MCMC explores pedigrees and mating parameters simultaneously. To explore the pedigree space, we make local modifications to the edges and the nodes in the graph using 12 reversible updates. The 12 updates can broadly be categorized into two groups. The first category of updates involves inserting or deleting edges to join or split pedigrees. The

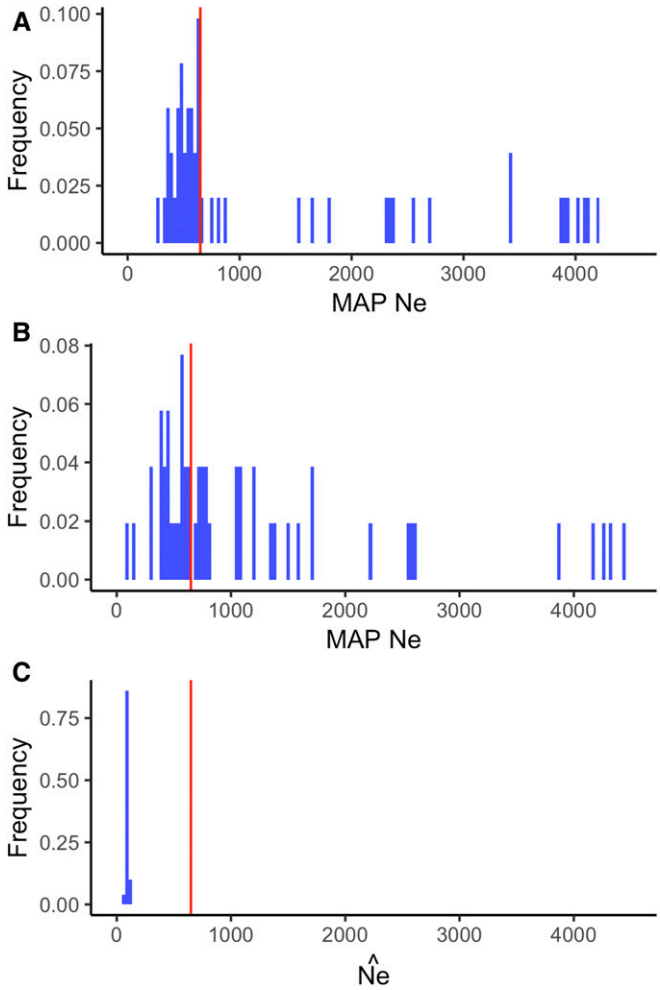

Figure 4 Distribution of the N_e estimates in Simulation B (*i.e.*, microsatellites). (A) Distribution of MAP N_e estimated from MCMC samples under two-generation inference. (B) Distribution of MAP N_e estimated from MCMC samples under one-generation inference. (C) Distribution of N_e estimate by COLONY under nonrandom mating.

second category modifies the pairwise relationship between two randomly chosen individuals, such as changing half-siblings to full-cousins, and vice versa. To explore the mating parameters, we use three different updates—one for each mating parameter—where we propose a new state by sampling the new parameter value from a normal distribution centered at the current value. A more detailed treatment of the updates is given in *Appendix*.

Here, we outline the MCMC algorithm. Let $Q = (H, \theta)$ denote the set of parameters we want to estimate (*i.e.*, pedigree and mating parameters).

1. Initialize pedigree H to be the one in which every individual is unrelated to each other. Initialize α by sampling from $U(\alpha_{min}, \alpha_{max})$, for some fixed α_{min} and α_{max}. Initialize β and N in a similar way. Compute and store Equation 5 for the current configuration.
2. Choose 1 of the 10 updates at random and generate a new configuration.

Table 4 Pairwise prediction accuracy for simulation B (microsatellites), aggregated over 50 independent datasets

		Predicted				
		FS	HS	UR	FC	HC
True	FS	96	22	7	0	0
	HS	2	31	81	0	0
	UR	0	23	60,054	0	0
	FC	1	8	445	0	0
	HC	0	0	480	0	0

Two-generation inference by MCMC.

3. If the new configuration is invalid, reject and go back to step 1. If it is valid, accept the new configuration with probability.

$$min\left(1, \frac{CL(H_{new})Pr(H|\theta_{new})Pr(Q_{old}|Q_{new})}{CL(H_{old})Pr(H|\theta_{old})Pr(Q_{new}|Q_{old})}\right).$$

4. Repeat steps 1–3 T times.

The total number of samples, T, was chosen to achieve a balance between convergence of the Markov chain and computational time. Since we want to keep samples only after the Markov chain has converged to the stationary distribution, we discarded the first B samples as burn-in. To check for convergence, we ran multiple independent MCMC chains and checked that all chains fluctuated in a similar, stable range of log-likelihood values. We note that this is only a proxy for checking convergence and there are other, albeit more involved, methods, such as checking the potential scale reduction factor for some specified quantity (Gelman and Rubin 1992). Furthermore, we keep only every tth sample to avoid storing correlated samples.

For both simulated and empirical datasets, which will be described next, we ran the MCMC for $T = 6 \times 10^6$ iterations with a burn-in period of $B = 4 \times 10^6$ iterations. The hyperprior for the mating parameters was set as follows: $\alpha \sim U(.1, 100)$, $\beta \sim U(1 \times 10^{-5}, .1)$, and $N \sim U(5, 5000)$. We also thinned the MCMC samples by keeping only every 50th sample.

Simulated data

We tested the performance of our method on simulated data. We simulated pedigrees up to two generations back in time using the mating model described in Prior with $\alpha = 15$, $\beta = 1e-4$, and $N = 1000$, which translates to $N_e = 650$ using the formula given in Gasbarra *et al.* (2005). The sample size, n, was 50.

We then simulated 10,000 independent single nucleotide polymorphic sites (SNPs) for each of the N founders in the pedigree, where the population allele frequency for each marker was sampled from the site frequency spectrum under neutrality. We assumed that the markers were spread evenly among 20 independent chromosomes of length 100 Mb, and assumed sequencing error rate of 0.01. To test the effect of marker type on our parameter inference, we also simulated

Table 5 Pairwise prediction accuracy for simulation B (microsatellites), aggregated over 50 independent datasets

		Predicted		
		FS	HS	UR
True	FS	91	22	12
	HS	2	25	87
	UR	1	23	60,053
	FC	0	3	451
	HC	0	0	480

One-generation inference by MCMC.

Table 6 Pairwise prediction accuracy for simulation B (microsatellites), aggregated over 50 independent datasets

		Predicted		
		FS	HS	UR
True	FS	102	22	1
	HS	2	92	22
	UR	3	1675	58,399
	FC	1	105	348
	HC	0	39	441

One-generation inference by COLONY.

20 microsatellites with 10 alleles of equal frequency per marker. Furthermore, we assumed that each marker was on an independent chromosome, with sequencing error rate of 0.01 and allele dropout rate of 0.05.

We then simulated the genotypes for the children in the pedigree by recombining parental haplotypes at rate 1.3e−8 per base pair per generation. We generated 50 independent datasets for both SNP and microsatellite simulations. For convenience, in subsequent sections, we refer to the simulations with SNPs as Simulation A, and those with microsatellites as Simulation B.

Bias correction

For real datasets, it is often unreasonable to assume that the sample does not contain relatives more distant than first cousins. To show the effect on the inference of pedigrees and N_e of having second cousins in the sample, we simulated a scenario where second cousins were present in the sample. The simulation parameters were identical to those of Simulation A, except for the number of generations under which the pedigrees were simulated. Instead of going back up to two generations back in time as in Simulation A—which generated relatives up to first cousins—here, we simulated pedigrees up to three generations back in time, which generated second cousins as well.

As we will discuss in *Results*, the second cousin relationships will often be classified as half first cousins (HC), the most distant relationship type our method is designed to estimate. Consequently, N_e will be biased downward due to the high frequency of HC in the estimated pedigrees, caused by the misclassification of second cousins as HC.

To correct the downward bias in N_e estimation, we took advantage of the fact that our method can still infer siblings with high accuracy. More specifically, we simulated pedigrees under various N_e to find a value that generated a number of siblings close to the one estimated by our method. Let $S_{IBD} = N_{FS} + .5N_{HS}$ be the summary statistic that measures the level of identical-by-descent (IBD) contributed by siblings in the sample, where N_{FS} and N_{HS} are the number of full siblings and half siblings, respectively; and denote $\hat{S}_{IBD}$ to be the statistic obtained from the MCMC inference on the sample. Let α_{MAP} and β_{MAP} be the MAP estimates of α and β, respectively, computed using the marginal posterior distributions obtained from the MCMC samples. We then simulated pedigrees going back up to one generation in time under α_{MAP}, β_{MAP}, and various values of N—which translate to different values of N_e—and computed S_{IBD} from the simulated pedigrees. We then chose the value of N_e that produced S_{IBD} that most closely matched $\hat{S}_{IBD}$.

Empirical data

We applied our method to reconstruct the previously unreported pedigree of house sparrows collected from an archipelago off the Helgeland coast of northern Norway (Lundregan *et al.* 2018). The individuals were genotyped using a custom Affymetrix 200K SNP array, with markers distributed across 29 of the chromosomes in the genome. Also provided were the location and year in which each individual was collected.

We used individuals from a single island (island 27) to avoid any potential substructure in the sample. Furthermore, we restricted our analysis to the individuals born in 2009 to ensure that all samples belonged in a single generation. We pruned the markers for LD using PLINK (Chang *et al.* 2015) at $r^2 = 0.05$ to get a set of independent, or loosely linked, markers. The filtering steps resulted in 79 individuals and 4519 SNPs. The likelihoods were computed by Albrechtsen *et al.* (2009) for linked markers.

Evaluation of method

We compared the performance of our method to that of COLONY (Jones and Wang 2010)—one of the most widely used pedigree reconstruction methods. We chose COLONY for several reasons. First, it supports full likelihood computation, which provides a gold standard to which we can compare our composite likelihood method. Second, it supports both SNPs and microsatellites data, allowing us to compare the performance of different marker types. Third, COLONY can estimate the short-term N_e based on the estimated frequency of siblings in the sample, which was shown to be more accurate than other methods of estimating N_e (Wang 2009).

Because the sample size in our simulations was much smaller than the population size, many pedigrees for the sample had similar likelihoods, making it difficult for both our method and COLONY to find the correct pedigree in its entirety. So we used pairwise prediction accuracy as a proxy for the accuracy of pedigree inference. In our method, we assigned pairwise relationship R to pair (i,j) if it had the

Table 7 Pairwise prediction accuracy for datasets containing second cousins (inference by MCMC), aggregated over 50 independent datasets

		Predicted				
		FS	HS	UR	FC	HC
True	FS	118	1	0	0	0
	HS	0	108	2	0	1
	UR	0	0	56,189	0	3
	FC	5	5	0	386	95
	HC	0	0	9	4	499
	2FC	0	0	523	2	1388
	2HC	0	0	1,482	0	430

highest posterior probability among all competing relationships. We approximated the posterior probability of R by counting the proportion of times pair (i,j) had relationship R in the MCMC samples. Similarly, we assigned relationship R to pair (i,j) in COLONY if it had the highest probability among all candidate relationships. Because the number of possible pedigrees is large, COLONY archives only the top w pedigrees with highest likelihoods. Suppose S is the set of indices for the pedigrees where (i,j) has relationship R. Then, the probability of R is estimated by

$$\frac{\sum_{k\in S} L_k}{\sum_{m=1}^{w} L_m},$$

where L_m is the likelihood of the mth pedigree.

Furthermore, since COLONY restricts its inference to pedigrees going back only one generation back in time (*i.e.*, siblings), we also limited our inference to the same scope when comparing the performance of our method to COLONY. The parameters used to run COLONY are detailed in Supplemental Material, File S1.

Data availability

Our software for pedigree inference is available for download at https://github.com/amyko/mcmcPed. Simulated data are available upon request. All supplemental files are available at FigShare. Supplemental material available at FigShare: https://doi.org/10.25386/genetics.8079953

Results

Simulated datasets

To illustrate some of the issues involved in estimating multigeneration pedigrees, we first turn our attention to an example from Simulation A. Figure 1 shows the two most likely local pedigrees involving three sampled individuals (shaded) and their estimated posterior probabilities. In the first pedigree, individual 3 forms a full first-cousin relationship with the other two individuals (1 and 2), as opposed to a half first-cousin relationship as in the second pedigree. Here, the true pedigree is shown by the first pedigree (Figure 1A), which had the highest posterior probability.

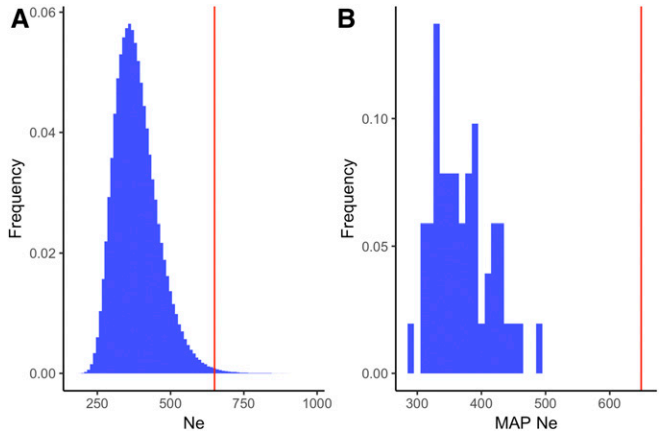

Figure 5 (A) Estimated posterior distribution of *Ne* from MCMC samples aggregated over 50 datasets, where the data contained second cousins. (B) Distribution of MAP N_e for the 50 datasets.

The uncertainty in the pedigree estimation, shown by the similar posterior probabilities of the two pedigrees (0.55 and 0.45), was consistent with the fact that the pairwise likelihood values were similar under different relationships. More specifically, individuals 1 and 3 had a higher likelihood of being full cousins than half cousins by ∼1 log likelihood unit. On the other hand, individuals 2 and 3 had a higher likelihood of being half cousins than full cousins by roughly the same amount. Based on pairwise likelihoods alone, individuals 1 and 3 would be classified as half cousins, and individuals 2 and 3 as full cousins. Piecing together such pairwise assignments, however, would not produce a valid pedigree. Such uncertainties in cousin inference were not uncommon: ∼20% of true cousin pairs in Simulation A had nonzero posterior probabilities for both full and half cousins.

Table 1 shows the pairwise prediction accuracy of MCMC for the 50 independent datasets in Simulation A, where the pairwise likelihoods were computed using the method by Albrechtsen *et al.* (2009). Full siblings, half siblings, and half cousins were classified correctly in almost all instances, whereas ∼7% of full cousin pairs were classified as half cousins. The rate of false detection of relatives was very low at ∼0.01%, where the unrelated pairs were estimated as half cousins.

Figure 2A shows the posterior distribution of N_e estimated from the MCMC samples aggregated over the 50 datasets in Simulation A. The mode of the posterior distribution was close to the true value, indicated by the red vertical line. Similarly, Figure 2B shows that the distribution of maximum *a posteriori* (MAP) N_e for the 50 datasets was concentrated around the true value. The three mating parameters that make up the components terms of N_e (*i.e.*, α, β, and N) showed high correlations among them. Figure S1 shows that high values of N tended to co-occur with low values of α for this simulation, which suggests that these parameters should not be estimated independently of each other, and that marginal point estimates of any of these parameters are likely to be misleading.

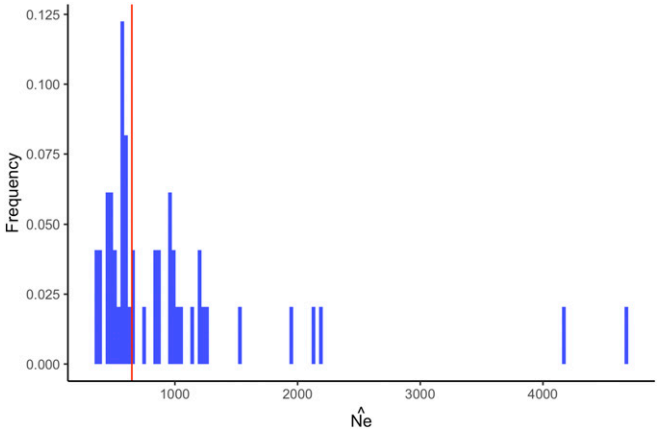

Figure 6 Distribution of the N_e estimates for the 50 datasets after bias correction. The red vertical line indicates the true value of N_e.

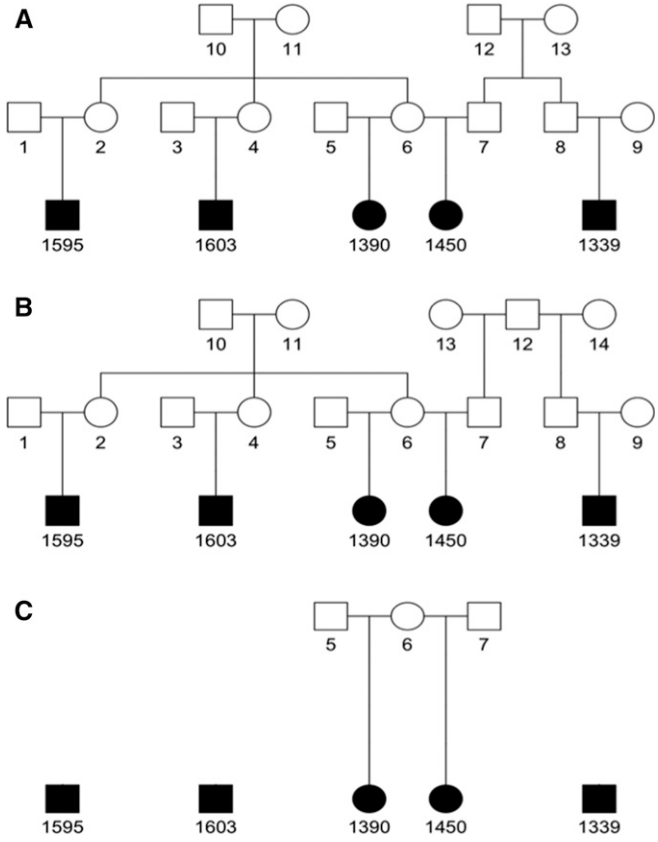

Figure 7 Estimated pedigrees of five sampled individuals in the sparrow dataset. (A) Pedigree with estimated posterior probability of 0.77. (B) Pedigree with estimated posterior probability of 0.23. (C) Most likely pedigree estimated by COLONY, but whose posterior probability was zero in our method.

Table 2 and Table 3 compare the performance of our method with that of COLONY. Since COLONY estimates up to sibling relationships only, we also restricted the inference of our method to the same scope. Furthermore, we computed the likelihoods using the method discussed in Weir *et al.* (2006), which assumes unlinked markers, an assumption that COLONY makes in its likelihood computation. Here, both our method and COLONY classified full siblings, half siblings, and unrelated pairs without error. Both methods also estimated all half cousin pairs to be unrelated. Furthermore, similar proportions of full cousin pairs were misclassified as half siblings by both methods: 22% by COLONY and 24% by our method. As shown in Figure 3, N_e was underestimated by both methods, which is consistent with the higher proportion of half siblings in the estimated pedigrees, caused by the misclassification of some full cousin pairs as half siblings.

File S2 shows the performance of our method on a simulated dataset with a different set of mating parameters (*i.e.*, $\alpha = 15, \beta = 0.1, N = 1000$), which translates to N_e of 958. The accuracy of pairwise relationship prediction and N_e estimation was similar to that of Simulation A.

Table 4 shows the pairwise prediction accuracy of MCMC for Simulation B (*i.e.*, microsatellites), where the likelihoods were computed using the method of Wang (2004). The accuracy rates were significantly lower than those in Simulation A (*i.e.*, 10,000 SNPs). About 77% of full siblings and 27% of half siblings were classified correctly, and virtually all cousin pairs were estimated to be unrelated. This is likely due to the prior, which puts higher probabilities on sparsely connected pedigrees, overwhelming the likelihoods that do not show strong evidence for individuals being related. The distribution of MAP N_e also had a much higher variance compared to that of Simulation A (Figure 4A).

Table 5 and Table 6 compare the performance of our method with that of COLONY for Simulation B. Again, we restricted the inference by our method to sibships to make a fair comparison with COLONY. Here, COLONY performed better than our method in correctly inferring full siblings and half siblings, but it also had a much higher false positive rate of 2.8% compared to 0.04% in our method. In fact, ~87% of the pairs estimated as half siblings by COLONY were actually unrelated. We note, however, that this problem may be addressed by adding an appropriate prior that is more conservative in half sibling assignments. Furthermore, due to the large number of unrelated pairs and cousins that were misclassified as half siblings, N_e was significantly underestimated by COLONY (Figure 4C).

For all the experiments, we checked the convergence of MCMC by studying the likelihood trace of multiple independent chains. As an illustration, we show an example of the log likelihood trace for the last 1 million iterations for a single experiment in Simulation A (Figure S2).

The running time for our method depends on many factors, such as the sample size, the underlying pedigree structure, and the maximum number of generation allowed in the pedigree inference. As an example, an MCMC run with 6 million iterations for a two-generation pedigree inference took ~36 sec on a laptop with 2.3 GHz Intel Core i5 processor for a single dataset in Simulation A, excluding the precomputation time for calculating the likelihoods. The precomputation time

Table 8 Comparison of pairwise relationship classification by MCMC and COLONY

		COLONY[a]		
		FS	HS	UR
MCMC[b]	FS	33	0	0
	HS	0	23	0
	UR	0	1	2909
	FC	0	15	37
	HC	1	4	57

[a]Inference was based on the full likelihood method, assuming independent markers.
[b]The likelihoods were computed by Albrechtsen *et al.* (2009) for linked markers and the inference allowed pedigrees going up to two generations back in time (*i.e.*, up to first cousins).

for the marginal and pairwise likelihoods by the HMM Albrechtsen *et al.* (2009) was ~75 sec.

Effects of presence of relatives beyond first cousins

Table 7 shows the prediction accuracy for a simulation scenario where second cousins were present in the sample. The accuracy rates were similar to those of Simulation A for relationships up to first cousins. However, ~73% of full second cousins (2FC) were classified as half first cousins (HC), the most distant relationship type our method is designed to estimate. Similarly, ~22% of half second cousins (2HC) were classified as HC. As expected, N_e was biased downward due to the high frequency of HC in the estimated pedigrees (Figure 5).

Figure 6 shows the distribution of the N_e estimates after correcting for bias as described in *Bias Correction*. Although the SE was higher than that of uncorrected estimates, the median of the distribution (657) was much closer to the true value (650) than before.

Sparrow dataset

We analyzed a subset of the house sparrow dataset sequenced by Lundregan *et al.* (2018). After the filtering steps described in *Empirical Data*, the sample consisted of 79 individuals and 4519 SNPs distributed across 29 autosomes. Here, we show an example of the inferred pedigrees by our method, and compare them to those estimated by COLONY.

Figure 7 shows the likely local pedigrees involving five individuals (shaded) in the sparrow dataset. The estimated posterior probabilities of the pedigrees shown in panels A and B were 0.77 and 0.23, respectively. The difference between the two pedigrees was the pairwise relationship between individuals 1339 and 1450, which was estimated to be full cousins in panel A and half cousins in panel B. Figure 7C shows the pedigree with the highest likelihood estimated by COLONY. This pedigree had posterior probability of zero in our method. We see that the half sibling relationship between individuals 1390 and 1450 was recovered by COLONY, but all cousin relationships that our method detected were estimated to be unrelated. Based on the simulation studies in *Simulated Datasets*, however, we expect the full first cousin relationships inferred by our method to be either true first cousins or, with considerably smaller probability, more distant relatives (*e.g.*, second cousins).

Table 8 compares the pairwise relationship classifications between our method and COLONY. Pairs that were classified as full siblings, half siblings, or unrelated by our method largely agreed with the classifications by COLONY. On the other hand, ~29% of pairs that were estimated to be full cousins by our method were estimated to be half siblings by COLONY, which is consistent with what was observed in the simulation studies in *Simulated Datasets*. Furthermore, most of the relationships that were inferred as half cousins by our method were classified as unrelated by COLONY.

Discussion

We have shown that, given enough marker information, our method is able to jointly estimate N_e and relationships up to first cousins accurately and efficiently. Unlike existing pedigree inference methods, our method not only allows estimation of pedigrees and N_e, but also provides an uncertainty measure on the estimates via posterior probabilities. Furthermore, our method provides a framework for incorporating different types of population models in the prior for the pedigrees, which can potentially allow us to estimate other population parameters, such as migration rates between subpopulations.

Our method also improves upon one of the most widely used pedigree reconstruction programs, COLONY, by estimating relationships beyond sibships. This not only expands the types of pedigrees we can infer, but also increases the accuracy of sibship inference. In particular, first cousins were often misclassified as half siblings if the estimation method did not allow inference of cousins. For example, ~44% of half siblings estimated by COLONY using 10,000 SNPs were actually first cousins (Table 3). Furthermore, we showed that N_e can be underestimated if the sample contains cousins but the pedigree inference is restricted to sibships only (Figure 3). By explicitly including first cousins in the inference, our method was able to infer half siblings with higher precision (Table 1), as well as estimate N_e more accurately (Figure 2). However, we note that the problem persists when the sample contains relatives more distant than first cousins. When datasets contained second cousins, for example, they were often estimated as half first cousins—the most distant relationship our method is designed to estimate—and, consequently, caused a downward bias in N_e estimates. Therefore, we must use caution in interpreting inferred half cousins, as the true relationship could be more distant, and use the simulation method discussed in *Effects of Presence of Relatives Beyond First Cousins* to correct for potential bias in N_e estimates.

We note that the performance of our method relies heavily on the accuracy of pairwise likelihoods. The accuracy of pairwise likelihoods depends on many factors, such as marker density, level of LD, sequencing error rates, and population allele frequency estimates. Ignoring the linkage between markers, in particular, significantly decreased the power to detect first cousins (File S3). Due to linkage, close relatives

such as first cousins are expected to share, with high probability, long IBD segments that are on the order of megabases in length, although the probability of IBD per marker is relatively low (Chapman and Thompson 2003). The presence of such long IBD segments should make detecting relatives quite easy even though identifying the exact relationship can be more difficult (Hill and White 2013). Treating the markers as independent, however, does not take advantage of the presence of long IBD segments, and, thus, decreases our ability to detect relatives (Table 2 and Table 3). Therefore, likelihood computation methods, such as that of Albrechtsen *et al.* (2009), that take into account the linkage information between markers should be used instead for detecting relatives, and, naturally, for pedigree inference as well.

Marker type and density also have a significant impact on the quality of pairwise likelihoods. We have seen that using 20 highly informative microsatellites performed worse than using 10,000 SNPs. The accuracy rates of COLONY (Table 6) suggest that the use of microsatellites to estimate sibships might be misguided in practice since first cousins can often be misclassified as half siblings in methods that do not explicitly model first cousins. Furthermore, microsatellites may not provide enough information to easily distinguish between full and half siblings (Table 4, Table 5, and Table 6). Also, 20 microsatellites with 10 alleles of equal frequency in our simulations is more generous than what is available in many real datasets, and the performance on less informative datasets is likely to be worse than what was shown in this study. We note that finding the best ways to address the various challenges in pairwise likelihood computation is an active area of research and requires further investigation.

There are limitations to our method that require further work. Our method does not support pedigrees that contain cycles, except those caused by full sibling relationships. More specifically, we do not consider pedigrees that are inbred or have complex, cyclic relationships such as double first cousins. A simulation study by Ko and Nielsen (2017) suggests that, in the presence of inbred individuals, the method will tend to estimate individuals to be genealogically closer than they actually are (*e.g.*, inbred first cousins estimated as half siblings). Furthermore, our method assumes that all samples belong in a single generation, which may not typically be true for many real datasets. This may be addressed by adding updates in the MCMC that allow sampled individuals to move between generations. Furthermore, our method does not yet scale up to sample sizes typical of GWAS as the number of pairwise comparisons still increases rapidly with sample size. One possible approach to address this issue is partitioning the sample into smaller sets using methods such as Manichaikul *et al.* (2010) and estimating the pedigrees for each smaller subset of individuals.

Overall, our method provides a way to jointly estimate pedigrees and N_e, and measure the uncertainty of the estimates in a computationally efficient way. Importantly, our method also provides a basic framework for estimating demographic parameters of the current population from pedigrees—analogous to population genetic methods based on coalescent trees—thus opening up new possibilities for learning about the demographic history of the recent past.

Literature Cited

Albrechtsen, A., T. Sand Korneliussen, I. Moltke, T. van Overseem Hansen, F. C. Nielsen *et al.*, 2009 Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet. Epidemiol. 33: 266–274. https://doi.org/10.1002/gepi.20378

Almudevar, A., 2003 A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theor. Popul. Biol. 63: 63–75. https://doi.org/10.1016/S0040-5809(02)00048-5

Almudevar, A., and E. C. Anderson, 2012 A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. Mol. Ecol. Resour. 12: 164–178. https://doi.org/10.1111/j. 1755-0998.2011.03061.x

Anderson, E. C., and T. C. Ng, 2016 Bayesian pedigree inference with small numbers of single nucleotide polymorphisms via a factor-graph representation. Theor. Popul. Biol. 107: 39–51. https://doi.org/10.1016/j.tpb.2015.09.005

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

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation plink: rising to the challenge of larger and richer datasets. Gigascience 4: 7. https://doi.org/10.1186/s13742-015-0047-8

Chapman, N., and E. Thompson, 2003 A model for the length of tracts of identity by descent in finite random mating populations. Theor. Popul. Biol. 64: 141–150. https://doi.org/10.1016/S0040-5809(03)00071-6

Cowell, R. G., 2009 Efficient maximum likelihood pedigree reconstruction. Theor. Popul. Biol. 76: 285–291. https://doi.org/10.1016/j.tpb.2009.09.002

Cowell, R. G., 2013 A simple greedy algorithm for reconstructing pedigrees. Theor. Popul. Biol. 83: 55–63. https://doi.org/10.1016/j.tpb.2012.11.002

Cussens, J., M. Bartlett, E. M. Jones, and N. A. Sheehan, 2013 Maximum likelihood pedigree reconstruction using integer linear programming. Genet. Epidemiol. 37: 69–83. https://doi.org/10.1002/gepi.21686

Elston, R. C., and J. Stewart, 1971 A general model for the genetic analysis of pedigree data. Hum. Hered. 21: 523–542. https://doi.org/10.1159/000152448

Eu-ahsunthornwattana, J., E. N. Miller, M. Fakiola, S. M. B. Jeronimo, J. M. Blackwell *et al.*, 2014 Comparison of methods to account for relatedness in genome-wide association studies with family-based data. PLoS Genet. 10: e1004445. https://doi.org/10.1371/journal.pgen.1004445

Gasbarra, D., M. J. Sillanpää, and E. Arjas, 2005 Backward simulation of ancestors of sampled individuals. Theor. Popul. Biol. 67: 75–83. https://doi.org/10.1016/j.tpb.2004.08.003

Gasbarra, D., M. Pirinen, M. J. Sillanpää, E. Salmela, and E. Arjas, 2007 Estimating genealogies from unlinked marker data: a Bayesian approach. Theor. Popul. Biol. 72: 305–322. https://doi.org/10.1016/j.tpb.2007.06.004

Gelman, A., and D. B. Rubin, 1992 Inference from iterative simulation using multiple sequences. Stat. Sci. 7: 457–472. https://doi.org/10.1214/ss/1177011136

Hadfield, J. D., D. S. Richardson, and T. Burke, 2006 Towards unbiased parentage assignment: combining genetic, behavioural

and spatial data in a Bayesian framework. Mol. Ecol. 15: 3715–3730. https://doi.org/10.1111/j.1365-294X.2006.03050.x
Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109. https://doi.org/10.1093/biomet/57.1.97
He, D., Z. Wang, B. Han, L. Parida, and E. Eskin, 2013 Iped: inheritance path-based pedigree reconstruction algorithm using genotype data. J. Comput. Biol. 20: 780–791. https://doi.org/10.1089/cmb.2013.0080
Hendricks, S., E. C. Anderson, T. Antao, L. Bernatchez, B. R. Forester *et al.*, 2018 Recent advances in conservation and population genomics data analysis. Evol. Appl. 11: 1197–1211. https://doi.org/10.1111/eva.12659
Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93: 47–64. https://doi.org/10.1017/S0016672310000480
Hill, W. G., and I. M. White, 2013 Identification of pedigree relationship from genome sharing. G3 (Bethesda) 3: 1553–1571. https://doi.org/10.1534/g3.113.007500
Jones, O. R., and J. Wang, 2010 Colony: a program for parentage and sibship inference from multilocus genotype data. Mol. Ecol. Resour. 10: 551–555. https://doi.org/10.1111/j.1755-0998.2009.02787.x
Kingman, J. F. C., 1982a Exchangeability and the evolution of large populations, pp. 97–112 in Exchange-Ability in Probability and Statistics, edited by G. Koch, and F. Spizzichino. North-Holland Publishing Company, Amsterdam.
Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. 19: 27–43. https://doi.org/10.2307/3213548
Kingman, J. F. C., 1982c The coalescent. Stochastic Process. Appl. 13: 235–248. https://doi.org/10.1016/0304-4149(82)90011-4
Kirkpatrick, B., S. C. Li, R. M. Karp, and E. Halperin, 2011 Pedigree reconstruction using identity by descent. J. Comput. Biol. 18: 1481–1493. https://doi.org/10.1089/cmb.2011.0156
Ko, A., and R. Nielsen, 2017 Composite likelihood method for inferring local pedigrees. PLoS Genet. 13: e1006963. https://doi.org/10.1371/journal.pgen.1006963
Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA 84: 2363–2367. https://doi.org/10.1073/pnas.84.8.2363
Lundregan, S. L., I. J. Hagen, J. Gohli, A. K. Niskanen, P. Kemppainen *et al.*, 2018 Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis. Mol. Ecol. 27: 3498–3514. https://doi.org/10.1111/mec.14811
Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale *et al.*, 2010 Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867–2873. https://doi.org/10.1093/bioinformatics/btq559
McPeek, M. S., and L. Sun, 2000 Statistical tests for detection of misspecified relationships by use of genome-screen data. Am. J. Hum. Genet. 66: 1076–1094. https://doi.org/10.1086/302800
Milligan, B. G., 2003 Maximum-likelihood estimation of relatedness. Genetics 163: 1153–1167.
Ott, J., Y. Kamatani, and M. Lathrop, 2011 Family-based designs for genome-wide association studies. Nat. Rev. Genet. 12: 465–474. https://doi.org/10.1038/nrg2989
Ramstetter, M. D., S. A. Shenoy, T. D. Dyer, D. M. Lehman, J. E. Curran *et al.*, 2018 Inferring identical-by-descent sharing of sample ancestors promotes high-resolution relative detection. Am. J. Hum. Genet. 103: 30–44. https://doi.org/10.1016/j.ajhg.2018.05.008
Riester, M., P. F. Stadler, and K. Klemm, 2009 Franz: reconstruction of wild multi-generation pedigrees. Bioinformatics 25: 2134–2139. https://doi.org/10.1093/bioinformatics/btp064
Smith, B. R., C. M. Herbinger, and H. R. Merry, 2001 Accurate partition of individuals into full-sib families from genetic data without parental information. Genetics 158: 1329–1338.
Staples, J., D. Qiao, M. H. Cho, E. K. Silverman, D. A. Nickerson *et al.*, 2014 Primus: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am. J. Hum. Genet. 95: 553–564. https://doi.org/10.1016/j.ajhg.2014.10.005
Staples, J., D. J. Witherspoon, L. B. Jorde, D. A. Nickerson, J. E. Below *et al.*, 2016 Padre: pedigree-aware distant-relationship estimation. Am. J. Hum. Genet. 99: 154–162. https://doi.org/10.1016/j.ajhg.2016.05.020
Steel, M., and J. Hein, 2006 Reconstructing pedigrees: a combinatorial perspective. J. Theor. Biol. 240: 360–367. https://doi.org/10.1016/j.jtbi.2005.09.026
Sun, L., and A. Dimitromanolakis, 2014 Prest-plus identifies pedigree errors and cryptic relatedness in the gaw18 sample using genome-wide SNP data. BMC Proc. 8: S23. https://doi.org/10.1186/1753-6561-8-S1-S23
Sun, L., M. Abney, and M. S. McPeek, 2001 Detection of misspecified relationships in inbred and outbred pedigrees. Genet. Epidemiol. 21: S36–S41. https://doi.org/10.1002/gepi.2001.21.s1.s36
Thatte, B. D., and M. Steel, 2008 Reconstructing pedigrees: a stochastic perspective. J. Theor. Biol. 251: 440–449. https://doi.org/10.1016/j.jtbi.2007.12.004
Thomas, S. C., and W. G. Hill, 2000 Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics 155: 1961–1972.
Thompson, E., 1975 The estimation of pairwise relationships. Ann. Hum. Genet. 39: 173–188. https://doi.org/10.1111/j.1469-1809.1975.tb00120.x
Vinkhuyzen, A. A. E., N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, 2013 Estimation and partition of heritability in human populations using whole-genome analysis methods. Annu. Rev. Genet. 47: 75–95. https://doi.org/10.1146/annurev-genet-111212-133258
Voight, B. F., and J. K. Pritchard, 2005 Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1: e32. https://doi.org/10.1371/journal.pgen.0010032
Wakeley, J., L. King, B. S. Low, and S. Ramachandran, 2012 Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190: 1433–1445. https://doi.org/10.1534/genetics.111.135574
Wakeley, J., L. King, and P. R. Wilton, 2016 Effects of the population pedigree on genetic signatures of historical demographic events. Proc. Natl. Acad. Sci. USA 113: 7994–8001. https://doi.org/10.1073/pnas.1601080113
Wang, J., 2009 A new method for estimating effective population sizes from a single sample of multilocus genotypes. Mol. Ecol. 18: 2148–2164. https://doi.org/10.1111/j.1365-294X.2009.04175.x
Wang, J., 2012 Computationally efficient sibship and parentage assignment from multilocus marker data. Genetics 191: 183–194. https://doi.org/10.1534/genetics.111.138149
Wang, J., and A. W. Santure, 2009 Parentage and sibship inference from multilocus genotype data under polygamy. Genetics 181: 1579–1594. https://doi.org/10.1534/genetics.108.100214
Wang, J., E. Santiago, and A. Caballero, 2016 Prediction and estimation of effective population size. Heredity 117: 193–206. https://doi.org/10.1038/hdy.2016.43
Wang, J. L., 2004 Sibship reconstruction from genetic data with typing errors. Genetics 166: 1963–1979. https://doi.org/10.1534/genetics.166.4.1963
Weir, B. S., A. D. Anderson, and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nat. Rev. Genet. 7: 771–780. https://doi.org/10.1038/nrg1960

Communicating editor: G. Coop

Appendix

Pedigree as a Graph

In our method, we represent pedigrees as an undirected graph, where each node represents an individual and an edge corresponds to a parent-offspring relationship. A node has the following features: sex, sample status, list of children, and list of parents. Sex refers to the sex of the node; sample status indicates whether the node was sampled and thus has genetic data associated with it; depth is the generation to which it belongs (*e.g.* most recent generation has depth 0); list of children contains the node's children nodes, if any; list of parents contains the node's parent nodes.

A valid pedigree is a graph that satisfies all of the following conditions:

1. Each individual, or node, has 0, 1, or 2 parent nodes. If it has 2 parents, then one parent must be female and the other male.
2. Graphs that contain cycles are not allowed, except for cycles caused by full sibling relationships. For example, any inbred pedigrees or non-full-sibling, cyclic relationships such as double first cousins are invalid.
3. If a node has two parents, the parents belong in the same generation. That is, the generations are nonoverlapping.
4. The height of the graph is ≤ 3. That is, the maximum number of generations in the pedigree we consider is 3, which includes up to first-cousin relationships.
5. All sampled individuals belong in the most recent generation.

We describe below the transitions between pedigree graphs used in the MCMC. For all of these transitions, or moves, if the resulting graph is invalid, we reject the move.

Transitions Between Pedigree Graphs

1. Link: join two pedigrees.
   (a) Choose a random pair of nodes i and j. Choose target depth k, drawn from a geometric distribution with $P = 0.5$. k is the depth at which i and j will share a common ancestor. If k is larger than the maximum depth of the pedigree, reject the move.
   (b) With equal probability, choose sex s (male or female) of the would-be common ancestor. Take a random path from node i up to the ancestor of sex s at depth k, choosing either the mother or the father at each step with equal probability. Do the same for node j.
   (c) At depth k, merge the two ancestors of i and j.

The reverse move is a combination of Cut and Split (see below).

2. Cut: detach a child and its subpedigree from a parent.
   (a) Choose node i at random. Choose sex s (male or female) with equal probability, which is the sex of the parent from which i will be cut. If the parent with sex s is not represented in the pedigree, reject the move.
   (b) Delete the edge between i and the parent.

   The reverse move is Link (see above).

3. Split: detach a set of children and its subpedigrees from a parent.
   (a) Choose node i at random. Choose a random set of i's children, where each set has an equal probability of being chosen.
   (b) Delete the edges between i and the set of children selected. Make a new parent node j and add edges between j and the set of children.

   The reverse move is Link (see above).

4. Switch Sex: switch the sex of a node.
   (a) Choose a random node i. Reject the move if i's sex cannot be changed (*i.e.*, i is sampled and its sex is fixed; or i has a spouse with fixed sex).
   (b) If i is female, switch its sex to male (and vice versa if i is male). If this sex change conflicts with the sex of other nodes, switch the sex of the other nodes as well. (*e.g.*, if i has a spouse, then the spouse must switch its sex as well).

   The reverse move is Switch Sex itself.

5. Full Sibs to Self: merge a pair of full sibling nodes into one node.
   (a) Choose node i at random. Choose at random node j among full siblings of i whose sex is the same as that of i. If no such node exists or if both i and j are sampled nodes, reject the move.
   (b) Merge i and j.

   The reverse move is Self to Full Sibs (see below).

6. Self to Full Sibs: split a node into a pair of full siblings.
   (a) Choose node i at random. Make a new node j, where the sex is the same as that of i.
   (b) Choose a random set of i's children, where each set has an equal probability of being chosen. Remove edges between the chosen children and i, and add edges between the children and j (*i.e.*, some of i's children are transferred to j).
   (c) Make i and j full siblings (*i.e.*, make them share the same mother and father).

   The reverse move is Full Sibs to Self (see above).

7. Self to Parents: split a single parent into a pair of parents (mother and father).
   (a) Choose node i at random. If i does not have exactly one parent, reject the move. Let p_1 be the parent of i.
   (b) Make a new node p_2 and set its sex to the opposite of p_1's sex. Set p_2 to be the other parent of i.
   (c) Choose a random set of p_1's parents, where each set has an equal probability of being chosen. Remove the edges between the chosen nodes and p_1, and add edges between the nodes and p_2 (*i.e.*, transfer some of p_1's parents to be p_2's parents).

   The reverse move is Parents to Self (see below).

8. Parents to Self: merge two parents into one node.
   (a) Choose node i at random. Reject if it does not have exactly 2 parents: p_1 and p_2.
   (b) Choose sex s (male or female) with equal probability.
   (c) Merge p_1 and p_2 into one node and set the sex to s.

   The reverse move is Self to Parents (see above).

9. MaternalHC to PaternalHC: change maternal half cousins into paternal half cousins.
   (a) Choose node i at random. If i does not have any maternal half cousins, reject the move. Choose at random node j among the maternal aunts and uncles of i.
   (b) Detach j and its children from i's maternal grandparent and attach them to i's paternal grandparent. In other words, i and the children of j are now paternal half cousins, not maternal.

   The reverse move is PaternalHC to MaternalHC (see below).

10. PaternalHC to MaternalHC: change paternal half cousins into maternal half cousins.
    (a) Similar to MaternalHC to PaternalHC above.

    The reverse move is MaternalHC to PaternalHC (see above).

11. MaternalHS to PaternalHS: change maternal half sibs into paternal half sibs, and vice versa.
    (a) Choose node i at random. Choose at random node j among the maternal half siblings of i. If no such node exists, reject the move.
    (b) Detach j and its children from i's mother and attach them to i's father. In other words, i and j are paternal half sibs, not maternal.The reverse move is PaternalHS to MaternalHS (see below).

12. PaternalHS to MaternalHS: change maternal half sibs into paternal half sibs, and vice versa.
    (a) Similar to MaternalHS to PaternalHS above.

    The reverse move is MaternalHS to PaternalHS (see above).

13. Update α
   (a) Given α_{current} and for some fixed variance σ_α^2, draw α_{new} from $N(\alpha_{\text{current}}, \sigma_\alpha^2)$.
   (b) If α_{new} does not lie between the prespecified bounds $[\alpha_{\min}, \alpha_{\max}]$, reject the move.The reverse move is Update α itself.

14. Update β
   (a) Given β_{current} and for some fixed variance σ_β^2, draw β_{new} from $N(\beta_{\text{current}}, \sigma_\beta^2)$.
   (b) If β_{new} does not lie between the prespecified bounds $[\beta_{\min}, \beta_{\max}]$, reject the move.

   The reverse move is Update β itself.

15. Update N
   (a) Given N_{current} and for some fixed variance σ_N^2, draw N_{new} from $N(N_{\text{current}}, \sigma_N^2)$. Round N_{current} to be an integer.
   (b) If N_{new} does not lie between the prespecified bounds $[N_{\min}, N_{\max}]$, reject the move.

   The reverse move is Update N itself.