Molecular Ecology (2006) **15**, 3715–3730 doi: 10.1111/j.1365-294X.2006.03050.x

Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework

J. D. HADFIELD*, D. S. RICHARDSON† and T. BURKE*
**Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK, †Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK*

Abstract

Inferring the parentage of a sample of individuals is often a prerequisite for many types of analysis in molecular ecology, evolutionary biology and quantitative genetics. In all but a few cases, the method of parentage assignment is divorced from the methods used to estimate the parameters of primary interest, such as mate choice or heritability. Here we present a Bayesian approach that simultaneously estimates the parentage of a sample of individuals and a wide range of population-level parameters in which we are interested. We show that joint estimation of parentage and population-level parameters increases the power of parentage assignment, reduces bias in parameter estimation, and accurately evaluates uncertainty in both. We illustrate the method by analysing a number of simulated test data sets, and through a re-analysis of parentage in the Seychelles warbler, *Acrocephalus sechellensis*. A combination of behavioural, spatial and genetic data are used in the analyses and, importantly, the method does not require strong prior information about the relationship between nongenetic data and parentage.

Keywords: maternity, MCMC, microsatellite, parentage analysis, paternity, Seychelles warbler

Received 6 November 2005; revision received 17 February 2006; accepted 17 March 2006

Introduction

In many ecological and evolutionary studies the pedigree structure of a sample of individuals is one of the most important pieces of information. Without it, many of the insights we have gained into the nature of natural selection, sexual selection, and the genetic basis of phenotypic variation would not have been possible. However, the pedigree structure of a sample, unlike many other pieces of information, is not normally directly observed and has to be inferred (reviewed in Blouin 2003). Here we focus on the methods, commonly referred to as parentage analyses, that are used to infer pedigree structure under the simplifying assumption that individuals are either drawn from an unrelated parental generation or from their progeny (reviewed in Jones & Ardren 2003).

Despite this simplifying assumption, the number of possible pedigrees may be very large, even when the number of individuals sampled from parental and offspring generations is relatively small (Almudevar 2003). Consequently, parentage analysis usually proceeds in two steps. The first step involves a mixture of demographic, sexual and behavioural data, logic and heuristics to restrict the number of possible pedigrees considered. Then, the second step evaluates the relative likelihood of those pedigrees considered possible, using genetic data to estimate the distribution of genotypes in the parental generation, and the probability of offspring genotypes conditional on the parental genotypes (Thompson 1976a). In reality, however, the ultimate goal of most studies is to estimate some population-level parameter, and the pedigree itself, although necessary, is nothing more than a nuisance parameter (Rosenberg & Nordborg 2002; Jones & Ardren 2003). Consequently, most parentage analyses are followed by a third step in which parameters of interest, such as heritability, mate choice and fecundity, are estimated from the pedigree information.

To simplify the discussion we have classified the large number of methods currently available into three classes: categorical allocation, fractional allocation, and full probability models (see Jones & Ardren 2003). These classes are

Correspondence: J. Hadfield, Fax: +44 (0)114 2220002;
E-mail: j.hadfield@shef.ac.uk

neither exhaustive nor mutually exclusive, and we have chosen them to best illustrate the range of situations that impact on the parameters estimated during step three (see information boxes). Accordingly, the metric by which we choose to compare methods is their ability to estimate the parameter of interest without systematic bias and with a high and quantifiable level of precision.

In categorical and fractional allocation models the parentage analysis (steps 1 and 2) and the population-level parameters subsequently estimated (step 3) are divorced from each other, and are done sequentially. This leads to the necessary decision of what information from step 2 should be used in step 3. In categorical allocation, the pedigree structure with the most support, the maximum-likelihood estimate in many cases, is used when estimating population-level parameters. This has the disadvantage that the most likely pedigree is assumed to be true, and any uncertainty regarding its structure is not taken into account when estimating the population-level parameters (Thomas 2005). Parameters estimated in this way are therefore likely to have inflated type I error rates, and to reduce this problem inferences are often limited to parts of the pedigree that have high support.

Fractional allocation models, on the other hand, do not use a single summary statistic at step 3, but estimate the parameters of interest from the complete probability distribution of the pedigree. In this way, fractional allocation models take into account uncertainty in the parameters that arises during step 2. For this reason, conclusions drawn from fractional allocation models will be more robust than conclusions drawn from categorical allocation models when genetic data are not highly informative, or when the number of pedigrees deemed possible at step 2 is large (Devlin *et al*. 1988).

Both fractional and categorical allocation models suffer from the fact that steps 2 and 3 are divorced from each other, and without modification tend to produce parameter estimates that are biased towards those that would be observed under random mating. For example, a common aim of many parentage analyses is to estimate the level of extra-pair paternity (EPP) in socially monogamous systems (Griffith *et al*. 2002). In these systems, naively treating all potential fathers as equally likely will lead to large upward biases in EPP rates if the true father is more likely to be the social father. Three solutions to this problem have been put forward.

Of great practical utility, but with poorly understood statistical properties, are the various heuristic methods for reducing the number of considered pedigrees at step 1. In the context of the previous example, a common procedure is to only consider extra-pair males as fathers when the social father mismatches at two or more loci. The rationale behind this approach stems from the belief that the probability of an offspring being sired by their social father is greater than the probability of being sired by an extra-pair father. Unfortunately, the strength of this belief, and its impact on parameter estimation, is hard to quantify and in the case of high genotyping error rates may actually be misguided.

The beliefs that motivate these heuristic methods can be formalized by treating them as prior information in a Bayesian analysis, as used by Neff *et al*. (2001). However, steps 2 and 3 still remain essentially divorced, and the authors themselves note the sensitivity of their analyses to this prior information. The third solution, which we refer to as full probability models, is to estimate parentage and the parameters of interest simultaneously, either in a Bayesian or maximum-likelihood framework (Roeder *et al*. 1989; Smouse & Meagher 1994; Burczyk *et al*. 1996; Smouse *et al*. 1999; Burland *et al*. 2001; Emery *et al*. 2001; Nielsen *et al*. 2001). Under these models the parameters that are typically estimated in step 3 are used to inform the parentage assignment in step 2. To illustrate, consider a sample of 20 potential fathers and 20 offspring, where each male is the social father of one offspring. For the first 19 offspring, the genetic data give very high support to the fact that each offspring's true father is their social father. The genetic data for the 20th offspring, on the other hand, give equal support to the social father and a single extra-pair male. When steps 2 and 3 are divorced from each other, the support for these two males remains equal. However, under a full probability model, the support given to the social male would be greater because the previous 19 offspring provide information that the social father is inherently more likely to be the true father. Using this information, unbiased estimates of the frequency of EPP would be obtained without the need for highly informative prior information.

Here, we develop a class of full probability models that simultaneously estimate parentage and a wide range of population-level parameters using a Markov chain Monte Carlo (MCMC) approach. We illustrate the method using data collected on the Seychelles warbler (*Acrocephalus sechellensis*), a system in which parentage analysis is particularly difficult (Richardson *et al*. 2001). Using simulated test data sets, we show that the method estimates the posterior distribution of both parentage and population-level parameters with accuracy when the assumptions of the model are met. We then fit these models to real data that have been the focus of a previous heuristic categorical parentage analysis (Richardson *et al*. 2001). We emphasize throughout a model-building approach to parentage analysis.

Materials and methods

The study population

The entire population of Seychelles warblers on Cousin Island was monitored during the peak breeding season in 1999 (see Richardson *et al*. 2001 for further details). During

this period the population consisted of 260 adult (> 6 months old) birds that had been individually colour ringed and blood sampled. The population consisted of 136 females and 124 males. In addition, blood samples were taken from 59 offspring produced during the season. All offspring and most adult birds could be unambiguously assigned to one of 104 territories based on observational data. The 59 offspring were distributed over 48 territories, with one territory having three offspring and nine having two. Of the 48 territories on which offspring were produced, 35 contained only a single adult female, 11 had two, and a single territory had three. Thirty-nine of the territories had a single territorial male, and nine had two. Territorial birds were categorized into dominants and subordinates based on behavioural data, with a single dominant bird of each sex present on each territory. The data presented have been the focus of a previous parentage analysis using a heuristic categorical approach (Richardson *et al*. 2001). In this earlier study the parentage of each offspring was reconstructed sequentially using biologically motivated logic. First, maternity was assigned by considering within-group females only. Following the maternity analysis, paternity was then assigned assuming the assigned mother was correct. Initially, paternity was assigned by considering within-group males only. For those offspring for which within-group males were excluded, extra-group males were then considered.

Microsatellite data

DNA extraction and genotyping were completed using a suite of 14 microsatellite loci developed in the Seychelles warbler, using the methodology described in detail in Richardson *et al*. (2000, 2001).

The model

Following Emery *et al*. (2001), random variables, $\mathbf{a}_f$, $\mathbf{a}_m$ and $\mathbf{a}_o$ denote candidate fathers, candidate mothers and offspring, with $\mathbf{a}_f^{(i)}$ and $\mathbf{a}_m^{(i)}$ denoting the proposed father and mother of offspring $\mathbf{a}_o^{(i)}$, respectively. Elements of $\mathbf{a}_o$ are unique and labelled 1 to n_o. Elements of $\mathbf{a}_f$ and $\mathbf{a}_m$ are not necessarily unique (if parents have multiple offspring) and are drawn from all potential fathers, labelled 1 to n_f, and all potential mothers, labelled 1 to n_m, respectively.

The matrix $\mathbf{P}^{(i)}$, of dimension $n_m \times n_f$, has all elements equal to zero except $\mathbf{P}^{(i)}_{\mathbf{a}_f^{(i)}, \mathbf{a}_m^{(i)}}$, which is equal to one. $\mathbf{P}^{(i)}$ follows a multinomial distribution with a single trial and $n_m \times n_f$ mutually exclusive and exhaustive outcomes:

$$\Pr(\mathbf{P}^{(i)} \mid \mathbf{A}^{(i)}) \sim \text{multin}(1, \mathbf{A}^{(i)}), \qquad \text{(eqn 1)}$$

where $\mathbf{A}^{(i)}$ is a matrix of equal dimension to $\mathbf{P}^{(i)}$ with element $\mathbf{A}^{(i)}_{j,k}$ representing the joint probability ($\mathbf{a}_m^{(i)} = j$, $\mathbf{a}_f^{(i)} = k$). The elements of $\mathbf{A}$ must lie between 0 and 1, since they are probabilities, and they must sum to 1, as each offspring must have a set of parents.

We model the set of multinomial probabilities (i.e. $\mathbf{A}^{(i)}$) as dependent on both genetic and nongenetic data. In the present context, we model the effect of three nongenetic variables, the territory and social status of potential mothers, and the distance between potential fathers and offspring. We wish to stress that the model is a generalized log-linear model (Smouse *et al*. 1999), and that other variables and their interactions could be included (e.g. sexual ornamentation, parental feeding, or a heritable trait, etc.).

We assume throughout that females can only gain maternity on their own territories, and so element $\mathbf{A}^{(i)}_{j,k}$ is equal to zero when the territory of the offspring is not the same as that of the potential mother j. In reality, the dimension of $\mathbf{P}^{(i)}$ is reduced to accommodate this, but we leave the notation with all mothers included as the assumption of no extra-group maternity can be relaxed.

The social status of each female is denoted by the vector $\mathbf{s}$, where 1 indicates that she is dominant, and 0 that she is subordinate. This variable is associated with the unknown parameter θ which is the probability that dominant mothers gain maternity over subordinate mothers.

The vector $\mathbf{d}^{(i)}$ denotes the distance between the centre of each male's territory and the territory on which offspring i was born. The probability of a father gaining paternity is assumed to follow an exponential function at a rate equal to the unknown parameter λ. When $\lambda > 0$ males close to the offspring are more likely to be the father and when $\lambda < 0$ distant males are more likely. If $\lambda = 0$ paternity is independent of distance.

The matrices $\mathbf{O}$, $\mathbf{F}$ and $\mathbf{M}$ represent the multilocus genotypes of offspring, potential fathers, and potential mothers, respectively. Rows index loci, of which there are L, and columns index individuals.

The joint probability that ($\mathbf{a}_m^{(i)} = j$, $\mathbf{a}_f^{(i)} = k$) is then proportional to the product of three probabilities: the probability that the jth father gains paternity given his distance from the ith offspring, the probability that the kth mother gains maternity given her social status, and the probability of offspring i's genotype conditional on both parental genotypes:

$$\mathbf{A}^{(i)}_{j,k} \propto \frac{\theta^{\mathbf{s}_j}(1-\theta)^{1-\mathbf{s}_j}}{\sum_{m=1}^{n_m}\theta^{\mathbf{s}_m}(1-\theta)^{1-\mathbf{s}_m}} \frac{e^{-\mathbf{d}_k^{(i)}\lambda}}{\sum_{f=1}^{n_f} e^{-\mathbf{d}_f^{(i)}\lambda}} \prod_{l=1}^{L} \frac{\Pr(\mathbf{O}_{l,i} \mid \mathbf{F}_{l,j}, \mathbf{M}_{l,k})}{\sum_{m=1}^{n_m}\sum_{f=1}^{n_f}\Pr(\mathbf{O}_{l,i} \mid \mathbf{F}_{l,f}, \mathbf{M}_{l,m})}. \qquad \text{(eqn 2)}$$

The model can be parameterized as a multinomial log-linear model (Gelman *et al*. 2004, p. 430) with $\mathbf{a}_m^{(i)} = 1$ and $\mathbf{a}_f^{(i)} = 1$ specified as baseline parents:

$$\begin{aligned}\eta_{j,k}^{(i)} &= \log(\mathbf{A}_{j,k}^{(i)}/\mathbf{A}_{1,1}^{(i)}) \\ &= \log\left(\frac{\theta^{s_j}(1-\theta)^{1-s_j}}{\theta^{s_1}(1-\theta)^{1-s_1}}\right) + \log\left(\frac{e^{-d_{ik}\lambda}}{e^{-d_{i1}\lambda}}\right) \\ &\quad + \sum_{l=1}^{L}\log\left(\frac{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,j},\mathbf{M}_{l,k})}{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,1},\mathbf{M}_{l,1})}\right) \\ &= \log\left(\frac{\theta^{s_j}(1-\theta)^{1-s_j}}{\theta^{s_1}(1-\theta)^{1-s_1}}\right) + \lambda(d_{i1}-d_{ik}) \\ &\quad + \sum_{l=1}^{L}\log\left(\frac{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,j},\mathbf{M}_{l,k})}{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,1},\mathbf{M}_{l,1})}\right).\end{aligned} \tag{eqn 3}$$

Ordering potential mothers, such that $s_1 = 1$, and that the indicator variables $\delta_j = 1$ when $s_j = 0$,

$$\begin{aligned}\eta_{j,k}^{(i)} &= \delta_j \log\left(\frac{1-\theta}{\theta}\right) + \lambda(d_{i1}-d_{ik}) \\ &\quad + \sum_{l=1}^{L}\log\left(\frac{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,j},\mathbf{M}_{l,k})}{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,1},\mathbf{M}_{l,1})}\right),\end{aligned} \tag{eqn 4}$$

which can be represented by the design matrix $\mathbf{X}$ and the vector of parameters β:

$$\eta_{j,k}^{(i)} = \mathbf{X}_{j,k}^{(i)}\beta, \tag{eqn 5}$$

where

$$\beta = \begin{bmatrix}\text{logit}(\theta) \\ \lambda \\ 1\end{bmatrix} \tag{eqn 6}$$

and

$$\begin{aligned}&\mathbf{X}_{j,1}^{(i)} = \delta_j \quad \mathbf{X}_{j,2}^{(i)} = d_{i1} - d_{ik} \\ &\mathbf{X}_{j,3}^{(i)} = \sum_{l=1}^{L}\log\left(\frac{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,j},\mathbf{M}_{l,k})}{p(\mathbf{O}_{l,i}\,|\,\mathbf{F}_{l,1},\mathbf{M}_{l,1})}\right).\end{aligned} \tag{eqn 7}$$

Making the rather strong assumption that the parentage of each offspring is independent, conditional on $\mathbf{X}\beta$, the likelihood of the vectors $\mathbf{a}_f$ and $\mathbf{a}_m$ is the product of n_o multinomial likelihoods:

$$\Pr(\mathbf{P}\,|\,\mathbf{X}\beta) = \prod_{i=1}^{n_o}\prod_{j=1}^{n_f}\prod_{k=1}^{n_m}\left(\frac{e^{\eta_{j,k}^{(i)}}}{\sum_{m=1}^{n_m}\sum_{f=1}^{n_f}e^{\eta_{m,f}^{(i)}}}\right)^{\mathbf{P}_{j,k}^{(i)}} \tag{eqn 8}$$

Implementation

We approximate the joint probability of the unknown parameters given the data, $\Pr(\mathbf{P}, \theta, \lambda\,|\,\mathbf{X})$, through a combination of Gibbs sampling and Metropolis–Hastings updates, with the following sequences of steps:

Step 1

Initialize the chain at time t by selecting starting candidates for all unknown variables, denoted $\mathbf{a}_f^t$, $\mathbf{a}_m^t$, λ^t and θ^t.

Step 2

Generate a new set of candidate parents, $\mathbf{a}_f^{t+1}$ and $\mathbf{a}_m^{t+1}$, conditional on $\mathbf{X}$, θ^t, λ^t using equation 8.

Step 3

Generate the candidate value $\theta^{t+1} \sim Un(0, 1)$.

Step 4

Using equation 8 and the prior distribution $\Pr(\theta) \sim \text{beta}(1, 1)$, evaluate the ratio:

$$a(\theta^t, \theta^{t+1}) = \frac{\Pr(\theta^{t+1}\,|\,\mathbf{P}^{t+1}, \lambda^t, \mathbf{X})}{\Pr(\theta^t\,|\,\mathbf{P}^{t+1}, \lambda^t, \mathbf{X})}. \tag{eqn 9}$$

Step 5

The new candidate, θ^{t+1}, is accepted with probability $\min[1, \mathbf{a}(\theta^t, \theta^{t+1})]$. If accepted, $\theta^{t+1} = \theta^{t+1}$, if not $\theta^{t+1} = \theta^t$.

Step 6

Generate the candidate value $\lambda^{t+1} \sim \text{norm}(\lambda^t, 0.02)$.

Step 7

Using equation 8 and the prior distribution $\Pr(\lambda) \sim \text{norm}(0, 100)$, evaluate the ratio:

$$a(\lambda^t, \lambda^{t+1}) = \frac{\Pr(\lambda^{t+1}\,|\,\mathbf{P}^{t+1}, \theta^{t+1}, \mathbf{X})}{\Pr(\lambda^t\,|\,\mathbf{P}^{t+1}, \theta^{t+1}, \mathbf{X})}. \tag{eqn 10}$$

Step 8

The new candidate, λ^{t+1}, is accepted with probability $\min[1, \mathbf{a}(\lambda^t, \lambda^{t+1})]$. If accepted, $\lambda^{t+1} = \lambda^{t+1}$, if not, $\lambda^{t+1} = \lambda^t$.

Step 9

$t = t + 1$ and repeat steps 1 to 9 until the chain converges.

The model was developed in C++ and R (R Development Core Team 2004) and made use of the Scythe statistical library (Martin *et al*. 2004). The code is available from the author for correspondence upon request.

Missing data and measurement error

A major assumption of the previous model is that the design matrix **X** is nonstochastic: the data are completely observed (i.e. there are no missing data), and the observed data are known with complete accuracy. In the presence of missing data, or measurement error, the predictor variables are best treated as stochastic, and their distributional form then needs to be considered explicitly. We deal with the problem of measurement error and missing data (including the special case of unsampled parents) using data augmentation. The idea behind data augmentation is to include additional steps in the Markov chain so that unobserved data can be simulated in a way that is consistent with the probability model. This simplifies the likelihood function, making the analyses tractable.

In the current data set, territory data are missing for three males, genotype data are missing for 32 loci across 12 individuals, and complete data records are missing for four unsampled individuals. In addition, genotypic data are rarely observed without error, and failing to accommodate genotyping error in likelihood calculations is known to have a large impact on parentage analyses (Marshall *et al*. 1998). Following Wang (2004), we acknowledge two broad classes of genotyping error: systematic (class I) and stochastic (class II). These two sources of error are a consequence of allelic dropout or other stochastic typing errors, respectively; they can have very different effects on pedigree reconstruction (Wang 2004) and can occur at very different frequencies (Gagneux *et al*. 1997).

Extending the method developed by Emery *et al*. (2001), the matrices of offspring, maternal and paternal genotypes (**O**, **M** and **F** of equation 2) are sampled conditional on the respective matrices of observed genotypes ($\mathbf{O}^{(obs)}$, $\mathbf{M}^{(obs)}$ and $\mathbf{F}^{(obs)}$), the pedigree structure (**P**), the allele frequencies (ω) in the base population and the genotyping error rates associated with class I and class II errors (ε_1 and ε_2, respectively):

$$\Pr(\mathbf{O}, \mathbf{M}, \mathbf{F} \mid \mathbf{P}, \mathbf{O}^{(obs)}, \mathbf{M}^{(obs)}, \mathbf{F}^{(obs)}, \varepsilon_1, \varepsilon_2, \omega) \propto \Pr(\mathbf{O}, \mathbf{M}, \mathbf{F} \mid \mathbf{P}, \omega)\Pr(\mathbf{O}^{(obs)}, \mathbf{M}^{(obs)}, \mathbf{F}^{(obs)} \mid \mathbf{O}, \mathbf{M}, \mathbf{F}, \varepsilon_1, \varepsilon_2) \quad \text{(eqn 11)}$$

An individual-by-individual Gibbs sampler was used to sample true genotypes according to equation 13, where the first term in the RHS of equation 13 is defined according to equation 6 in Sheehan (2000), and the second term according to equations 1 and 2 in Wang (2004). Point estimates for ε_1, ε_2 and ω were used (see Appendix). The error rates were estimated from samples that had been genotyped more than once, and the allele frequencies were estimated from the population as a whole, ignoring the pedigree structure. When samples have been genotyped more than once, then all sets of observed genotypes enter into equation 13, avoiding the problem of choosing which observed genotype is used to represent the genetic data.

The missing territory data were modelled in an analogous fashion. Each individual for which territory data were missing was assigned a territory during each iteration of the chain. The territory was sampled from a multinomial distribution, the parameters of which were modified according to a probability model consistent with equation 8. The number of males and females per territory were considered to be independent and Poisson distributed.

Test data

To evaluate the utility of the approach we created 300 test data sets. Each data set had the same structure as the real data set in that the first two columns of the design matrices were taken from the real data. For each data set, we generated the parents randomly from $[\mathbf{P}^{(i)} \mid \mathbf{X}^{(i)}\beta \sim \text{multin}(1, \mathbf{X}^{(i)}\beta)]$, with the third column of $\mathbf{X}^{(i)}$ set to 1's. Parental genotypes were then sampled at random, with frequencies equal to those estimated from the real data set under the assumption of Hardy–Weinberg and linkage equilibrium. Offspring genotypes were then sampled from parental genotypes following Mendelian inheritance (MacCluer *et al*. 1986). Genotyping errors were added at rates equal to those estimated from the real genotype data: $\varepsilon_1 = 0.029$ and $\varepsilon_2 = 0.01$. In all runs θ was set to 0.59 and λ to 0.239. The chain was run for 70 000 cycles with a burn-in of 30 000, and samples from the posterior distribution of β and **P** evaluated. In addition, β and **P** were evaluated using a categorical approach, where β was estimated using the parental vectors ($\mathbf{a}_m$, $\mathbf{a}_f$) for which the genetic data gave the most support. These estimates provided starting values for each chain in the Bayesian approach (i.e. $\mathbf{a}_f^t$, $\mathbf{a}_m^t$, λ^t and θ^t). Because categorical approaches often use subsets of the pedigree that are known with some level of confidence, we also estimated β using three sets of categorical assignments in which the confidence in the assignments was greater than 95% ($\alpha = 0.05$), 85% ($\alpha = 0.15$) and 75% ($\alpha = 0.25$), respectively. It should be noted that confidence was assessed at the level of individual assignments, rather than the population level (Marshall *et al*. 1998). When confidence at the individual level was set to be greater than 85%, this resulted in 5% of parents being misassigned at the population level.

So that a large number of test data sets could be analysed we assumed that all data were observed and that missing values were absent. We did however, include genotyping error, but rather than augmenting the design matrix with true genotypes, we modified the likelihood function to model this source of error:

$$\Pr(\mathbf{P}|\mathbf{O}^{(obs)}, \mathbf{M}^{(obs)}, \mathbf{F}^{(obs)}, \varepsilon_1, \varepsilon_2, \omega) \propto \sum_{\mathbf{G}} \Pr(\mathbf{P}|\mathrm{O}, \mathbf{M}, \mathbf{F}) \Pr(\mathbf{O}, \mathbf{M}, \mathbf{F}|\mathbf{O}^{(obs)}, \mathbf{M}^{(obs)}, \mathbf{F}^{(obs)}, \varepsilon_1, \varepsilon_2, \omega). \quad \text{(eqn 12)}$$

This method simply calculates the likelihood of parentage for all combinations of parental and offspring genotypes (loosely denoted as **G**), and then weights this probability by the likelihood of those combinations actually existing. The approach is standard (e.g. Marshall *et al*. 1998), but fails to acknowledge that **P** provides information on which individuals have been mistyped, and what their true genotypes are likely to be. Although this gives rise to logical inconsistencies, the practical consequences are unlikely to be large when genotyping errors are rare and the amount of missing genotype data small. Adopting the approach reduces the dimension of the Markov chain substantially, allowing us to analyse enough datasets to evaluate the statistical properties of the technique.

Real data

The model fitted to the test data was fitted to the real data, and samples from three separate chains were evaluated to assess convergence to the posterior distribution. The missing territory data and the true genotypes, however, were updated in the chains with the estimates of class I and II error rates fixed at $\varepsilon_1 = 0.029$ and $\varepsilon_2 = 0.01$ (see Appendix). These values were estimated from repeat sample genotypes and had 95% confidence intervals of 0.019–0.039 and 0.006–0.015, respectively. The sampling covariance between them was high with a correlation of −0.42. Chain 1 had the estimated parameters from a categorical approach as the starting parameterization. The starting parameterization of chain 2 was identical except θ° was set to 0.5 (maternity was independent of female status), and λ° was set to zero (paternity was independent of the distance between males and focal offspring). The starting parameterization of chain 3 was identical to that of chain 2 except that the second most likely parents from the categorical approach were used as the starting parameterization. Estimates were also made using the pedigree determined by the heuristic approach adopted by Richardson *et al*. (2001) using CERVUS (Marshall *et al*. 1998).

We also ran a separate Bayesian model in which extra-group paternity was modelled in addition to the simple distance-related measure. The unknown parameter associated with this (denoted ρ) can be thought of as the unknown probability that a within-group male gains paternity over an extra-group male. This parameter enters the model in the same way as θ, except the indicator variable now indicates whether the male is within-pair (1) or extra-pair (0). λ and ρ are expected to be strongly correlated since within-group males have, by definition, a distance of zero from the focal offspring. Inference about ρ is therefore dependent on λ, and is the increased (or reduced) chance of within-group males gaining paternity conditional on their chance given λ. Only in the special case when λ = 0 does ρ uniquely indicate the difference between within-group and extra-group males.

For each chain, 8000 values were sampled from the posterior distribution with a burn-in of 400 000 iterations and a thinning interval of 100. Convergence of the three chains to the same posterior distribution was assessed using the multivariate version of the potential scale reduction factor (Brooks & Gelman 1998). We used posterior predictive checking to assess whether the assumptions of the Bayesian model were more reasonable than models that ignore distance and maternal behaviour. As a measure of fit, we treated 60 randomly selected offspring genotypes as missing records and used the methodology described above to obtain predicted posterior distributions of the true genotypes. In the first analysis, λ and θ were estimated from the data, and in a second analysis λ and θ were constrained at 0 and 0.5, respectively. This second analysis is analogous to a fractional allocation model. The posterior distributions of the true genotypes were then compared with the observed genotypes. The posterior distributions of the true genotypes are only expected to converge on the observed genotypes when genotyping error is absent, however, with low genotyping error the model that fits the data best is likely to assign the highest density to the observed (but treated as missing) genotype.

Results

Test data

The Bayesian approach, on average, misassigned 4/59 parental pairs (range: 0–11) compared to 18 (range: 8–29) from a categorical analysis in which the most likely set of parents are assigned (Fig. 1). An average of 14/59 (0.02), 24/59 (0.05), and 30/59 (0.08) parentage assignments were retained in the categorical analyses where α was set to 0.05, 0.15, and 0.25, respectively. Values in parentheses refer to the proportion of those retained assignments that were misassigned, and can be thought of as the more commonly used population-wide confidence values (Marshall *et al*. 1998).

In all approaches, the mean estimates for θ (the probability that dominant mothers gain maternity over subordinate mothers) were not significantly different from the true underlying value ($F_{4,1495} = 0.29$, $P = 0.89$). However, the Bayesian approach and the categorical approach where the most likely set of parents were used, irrespective of confidence ($\alpha = 1$), showed greater precision, with the variance in the estimates being three times less than in the categorical analysis with α set to 0.05 (Fig. 2).

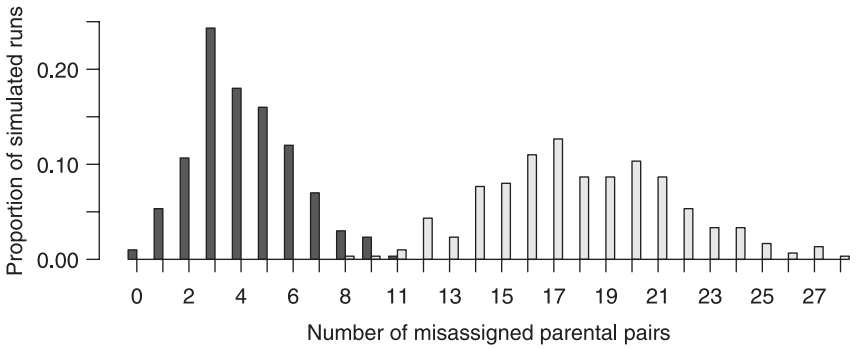

Fig. 1 The number of cases where the most likely parents were not the true parents based on a categorical analysis with α set to 1 (□), and the full probability model (■). The data presented are from 300 simulated test data sets.

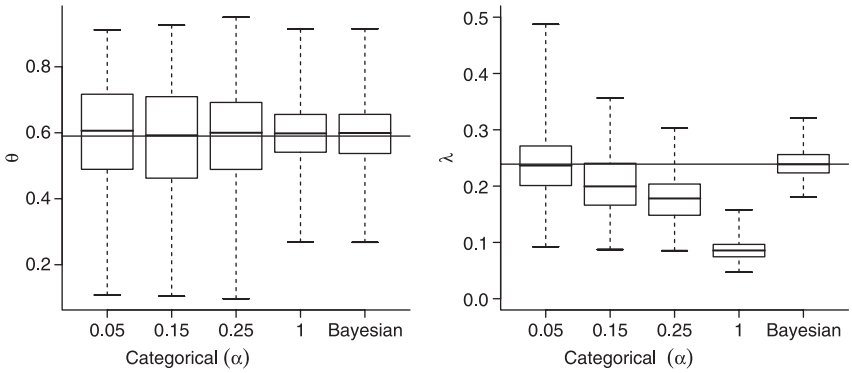

Fig. 2 The distribution of estimates of θ (the probability that dominant mothers gain maternity over subordinate mothers) and λ (the rate at which the chance of paternity drops with distance from an offspring) from the categorical and Bayesian approaches. In the categorical approach, the population-level parameters were estimated from a subset of the offspring that were assigned parents with varying levels of confidence (α). It should be remembered that these α values refer to confidence in individual assignments, rather than the population wide confidence usually used. Ninety-five per cent (95%) confidence at the population level corresponds to roughly α = 0.15 in this example. The whiskers of each box represent the minimum and maximum estimates, the boxes represent the interquartile range, and the median of the distribution is represented by the central line. The true underlying values of λ and θ are 0.239 and 0.59, respectively, for each of the 300 simulated test data sets. These values are represented by horizontal lines.

Estimates of λ (the rate at which the chance of paternity drops with distance from an offspring) were unbiased in the Bayesian approach, and the categorical approach when α was set to 0.05. However, the Bayesian approach was much more precise than the categorical approach, with the variance in the estimates being 0.0006 and 0.0035, respectively. Categorical analyses in which α was set to be greater than 0.05 showed less variance in the estimates, but the estimates were all downwardly biased (Fig. 2).

Table 1 Parameter estimates for θ, λ and EPP using different approaches

Method	θ	λ	% EPP
Categorical	0.49(0.26–0.73)	0.053(0.039–0.067)	61.8
Heuristic	0.66(0.40–0.86)	0.152(0.124–0.185)	38.2
Bayesian	0.59(0.34–0.81)	0.239(0.192–0.294)	44.1

θ is the probability that dominant mothers gain maternity over subordinate mothers and λ is the rate at which the chance of paternity drops with distance from an offspring. Confidence intervals of 95% are given in parentheses.

Real data

The Bayesian, heuristic (Richardson *et al*. 2001) and categorical approaches assigned maternity to the same mothers in all but two cases. For 10 out of 55 offspring, the most likely father using the Bayesian approach was not the father assigned using the heuristic approach with 75% confidence; in all but one case this was due to a different extra-pair male being assigned. For 28 out of 59 offspring, the most likely father using the Bayesian approach was not the father considered to be the most likely using the categorical approach. Dominant females had a slightly higher, but nonsignificant ($P = 0.23$), chance of gaining maternity than subordinates using the Bayesian approach (θ = 0.59). All approaches gave qualitatively the same answer (Table 1 and Fig. 3). The distance between an offspring and potential father was an important predictor of paternity,

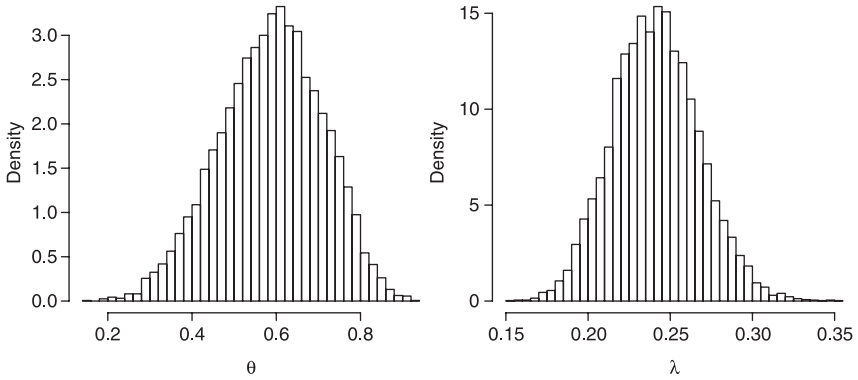

Fig. 3 Marginal posterior distributions of λ (the rate at which the chance of paternity drops with distance from an offspring) and θ (the probability that dominant mothers gain maternity over subordinate mothers) from a model without extra-pair/within-pair paternity modelled explicitly.

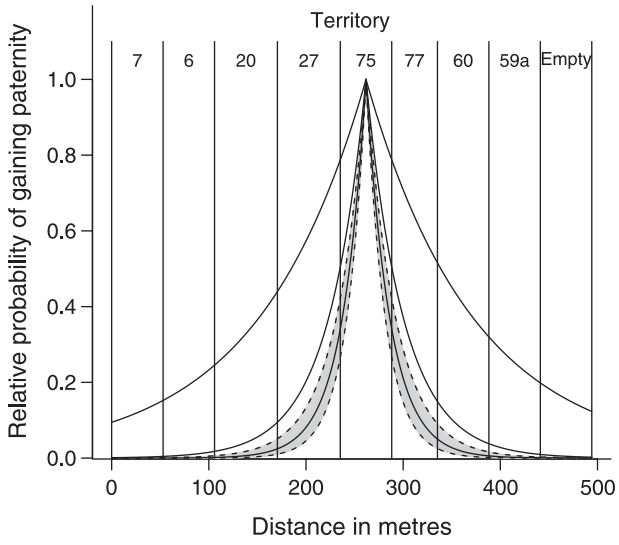

Fig. 4 An example of how the probability of paternity declines with distance for the three models. A transect is taken across the island, and the offspring hatched in territory 75. The outermost line is the decline in paternity estimated from a categorical analysis, the middle line is that estimated from a heuristic analysis, and the innermost shaded section is the estimate, and the 95% confidence intervals, from a Bayesian analysis.

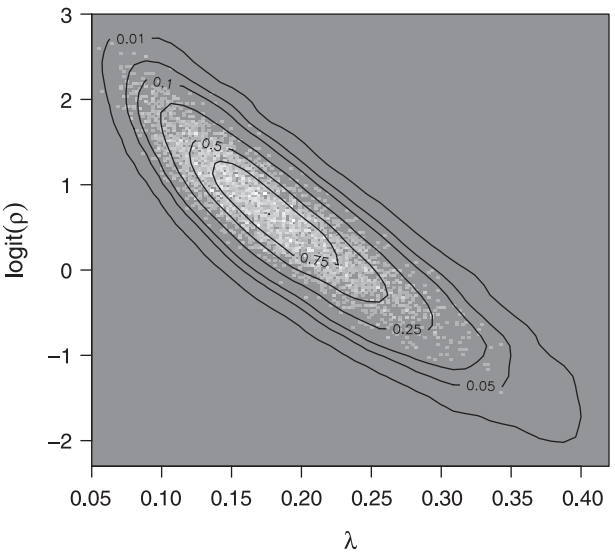

Fig. 5 Joint posterior distribution of λ (the rate at which the chance of paternity drops with distance from an offspring) and ρ (the probability that a within-group male gains paternity over an extra-group male), showing the strong sampling covariance between these two parameters. A value of zero for logit(ρ) indicates equal success by within-group and extra-group males at gaining paternity over and above the increase in paternity gained by being close to the offspring (λ). Positive values of logit(ρ) indicate that within-group males have greater success, and negative values indicate that extra-group males have greater success.

with a Bayesian estimate of $\lambda = 0.239$ (Table 1 and Fig. 3). The heuristic method gave estimates much closer to those that would be observed under random mating, and the 95% confidence intervals of the parameter estimated by the two methods did not overlap (Table 1 and Fig. 4). The categorical approach gave estimates of λ even closer to zero, with the 95% confidence intervals lying below those given by the heuristic method (Table 1 and Fig. 4).

Joint estimation of λ and ρ (within-group paternity) in the Bayesian approach showed that paternity could be adequately modelled by using distance alone, without the need to model extra-/within-pair paternity explicitly ($P = 0.15$) (see Fig. 5). However, the strong sampling covariance between the parameters, arising because within-group males have a distance of zero from the offspring, means that the power of this test was low. The marginal distributions of ρ and λ were 0.67 (0.30–0.89) and 0.190 (0.101–0.323), respectively. A value of $\rho = 0.5$ (0 on the logit scale, Fig. 5) indicates no difference between extra-pair and within-pair males at gaining paternity. Despite the posterior distribution of ρ indicating that this value is highly plausible, within-group males still have a higher chance of gaining paternity as the conditional distribution of λ when $\rho = 0.5$ does not overlap zero (Table 1). Since within-group males have a distance of zero from the offspring, the probability of gaining paternity is much higher than for males on other territories (Fig. 4).

The potential scale-reduction factor calculated across the three chains was close to 1 ($\hat{R} = 1.001$), indicating that the chains had converged on the same posterior distribution ($\hat{R}$ values < 1.1 are usually considered as evidence of convergence (Gelman *et al*. 2004)). The full-probability model was better at predicting missing genotype data than the fractional allocation model ($t = 2.054$, d.f. = 60, $P = 0.04$, paired *t*-test), indicating that the full-probability model gave a better fit to the data than the underlying model embodied by fractional and categorical methods.

Discussion

Here we present a general Bayesian method for incorporating genetic and nongenetic data into parentage analyses. We show that parentage, and parameters associated with parentage, estimated using this approach can be very different from those estimated using other techniques such as categorical allocation. We show that the full probability model developed gives unbiased estimates of population-level parameters when the assumptions of the model are met, and that estimates derived from categorical approaches may be severely biased.

When the number of potential parents per offspring is small, or when the genetic data are highly informative, the two approaches will converge on the same answer (see information boxes). Under these circumstances, parentage can be assigned with confidence using genetic data alone, and there will be little uncertainty in population-level parameters derived from uncertainty in parentage assignment. The real and test data sets analysed are a case in point. Because the mother of an offspring is considered a priori to be present in the same territory as the offspring, the number of potential mothers per offspring is low. The genetic data are therefore sufficient to identify maternity with confidence and the full probability method only slightly outperforms categorical analyses.

However, when genetic data are insufficient to identify parents with high certainty, categorical analyses perform poorly on two points. First, they fail to incorporate uncertainty in the population-level parameters that arises due to uncertainty in the parentage assignments. Second, they fail to use information that the population-level parameters provide on parentage assignment, resulting in bias and reduction in power (see information boxes). This is clearly exemplified in the real and test data sets. Previous work suggested that the frequency of extra-pair paternity in this population was high (Richardson *et al*. 2001), and consequently all adult males in the population were considered as potential fathers in the analysis. The large number of potential fathers means that the genetic data alone are not sufficient for assigning paternity with confidence, and many males have similar support regarding the paternity of some offspring. However, because the distance between male and offspring territories is highly informative as regards to paternity, this information can be used in the full probability model to increase the accuracy of parentage assignment. Moreover, because paternity and the spatial parameter are estimated simultaneously, the spatial parameter is not biased towards a value that would be observed under random mating, as is the case with categorical analyses where the spatial parameter is estimated post hoc (see Adams *et al*. 1992 and Fig. 4). The heuristic method employed by Richardson *et al*. (2001) only assigned extra-group males if the within-group male's LOD score was below a critical value (Marshall *et al*. 1998). With respect to estimating the rate of extra-pair paternity, this heuristic method is a great improvement over the naive categorical analysis; only a single offspring switched from being within-pair in the heuristic analysis to extra-pair in the Bayesian analysis. However, assigning paternity to specific extra-pair males differed markedly between the two approaches, with 10 out of 21 extra-pair offspring being assigned different fathers. These switches tended to be from fathers that were at a large distance from the offspring territory to those that were closer. A retrospective assessment of categorical assignment in red deer (*Cervus elaphus*) showed a similar pattern, with 63% of nonharem males that had been assigned at 80% confidence later found to have been misassigned (Slate *et al*. 2000).

In conclusion, full probability models offer the most powerful and accurate way of extracting information from pedigree information inferred using molecular markers. More popular techniques, where parameters of interest are estimated after parentage has been assigned, are inherently biased. This bias will inevitably be towards parameter values that would be observed under random mating, such as higher extra-pair paternity, lower heritability and weaker sexual selection on ornamental traits. Under some circumstances this bias may be small and of no practical concern; however, as demonstrated here, the biases may be large and biologically relevant. These problems have long been recognized by plant ecologists, and were even acknowledged in the original formulation of the fractional approach (Devlin *et al*. 1988). Consequently, modifications were soon developed to deal with the fractional approach's bias towards a Poisson distribution for the number of offspring per father (Roeder *et al*. 1989; Smouse & Meagher 1994). For identical reasons, estimates of pollen dispersal were found to be upwardly biased from fractional models, and less biased estimates were obtained by estimating parentage using genetic data and a preliminary estimate of pollen dispersal taken from a subset of the parentage assignments that had high genetic support (Adams *et al*. 1992). This approach was expanded on, and developed into a class of full probability models using maximum-likelihood and log-linear models (Burczyk *et al*. 1996, 2002; Smouse *et al*. 1999; Morgan & Conner 2001). Full-probability

models appeared much later in animal ecology and often appear to have been developed in ignorance of this earlier work (Burland *et al*. 2001; Emery *et al*. 2001; Jones & Clark 2003; but see Nielsen *et al*. 2001). Despite these later models lacking the generality of the log-linear model, they have extended the approach to cope with genotyping error (Emery *et al*. 2001), unsampled parents (Nielsen *et al*. 2001), and X-linked markers (Jones & Clark 2003). Despite these developments, categorical approaches still dominate molecular ecology, particularly in animal systems.

Here we introduce a Bayesian form of generalized linear modelling (McCullagh & Nelder 1989) where parentage and a wide range of population-level parameters can be estimated simultaneously. Any number of categorical and continuous variables can enter into the model, including interactions and quadratic terms (Smouse *et al*. 1999). The variables can relate to overall fecundity, or they can relate parents to specific offspring. These properties allow the strength and form of selection on a trait to be assessed (Smouse *et al*. 1999; Morgan & Conner 2001), and also the degree to which offspring resemble parents. The MCMC approach allows confidence in both population-level parameters and specific offspring–parent links to be evaluated directly, without having to make use of other simulation or resampling techniques. The model can be used whether maternity is known a priori, or not, and is able handle missing genetic and nongenetic data. We have also combined a more realistic model of genotyping error put forward by Wang (2004) with a method of estimating true genotypes and pedigree structure simultaneously (Emery *et al*. 2001). As such, the method can be used to identify mistyped loci and even assign a probability to possible genotypes.

Limitations

As in most parentage analyses, individuals in the parental generation are assumed to be unrelated such that the probabilities of maternal and paternal genotypes are independent. This is unlikely to be the case in many systems, and especially so in the example of the Seychelles warbler, where birds from the same territory may be quite closely related. Additional work is needed to assess how the model performs when this assumption is relaxed. However, joint assignment of paternity and maternity should reduce the bias that a related parental population induces, and may only be a major issue when the parental population contains birds that are related to the offspring as full siblings (Thompson 1976b). Future work dealing with the reconstruction of multigenerational pedigrees is planned.

The Seychelles warbler system is privileged in that the number of unsampled individuals in the population is low, and known with high accuracy. Consequently, the computational cost of updating the missing records is low, and the need to estimate the number of unsampled individuals is reduced. However, in many systems the number of unsampled individuals may be large and not known with a high degree of certainty (Nielsen *et al*. 2001). Under these circumstances the reversible jump MCMC method developed in Emery *et al*. (2001), and implemented in the software PARENTAGE, may be a useful improvement to the basic model we have developed here.

Variation in reproductive success, and mating systems such as monogamy, can all cause deviations from the Poisson process assumed in these models (Devlin *et al*. 1988; Neff *et al*. 2001). However, it is important to realize that the assumption of independent parentage is made after conditioning on the predictor variables, and deviations from the Poisson process can then be dealt with by including appropriate predictor variables. For example, nonindependence generated by monogamy can be adequately addressed by explicitly modelling within- and extra-pair parentage. When appropriate predictor variables are unavailable, we suggest hierarchical modelling of fertilities as an alternative. These fertilities could be defined at the level of the individual and/or the mated pair in order to model variation in reproductive success or alternative mating systems, respectively. A hierarchical approach to modelling fertilities may prove to be a more robust approach than that developed by Roeder *et al*. (1989), and may be less susceptible to overestimating variance in reproductive success (Morgan 1998).

Two Gibbs sampling strategies exist for sampling true genotypes according to equation 13: individual-by-individual Gibbs sampling and blocked Gibbs sampling where the genotypes of subsets of individuals are updated jointly (Sorensen & Gianola 2002). For highly polymorphic markers such as microsatellites, joint updating quickly becomes computationally unfeasible as family sizes increase, and an individual-by-individual Gibbs sampler was employed. However, it is well known from applications in segregation analysis that irreducibility of the Markov chain cannot be ensured by this method (Sheehan 2000). True irreducibility is unlikely to occur in the current application as the pedigree and the observed genotypes are not assumed to be fixed. However, transitions between likely configurations of **P**, **O**, **M** and **F** may be vanishingly small and mixing may be so slow that the chain may be considered irreducible in practical terms. The impact this will have on population-level parameter estimates requires further testing, and will depend on the structure of the data.

We have also used point estimates of genotyping error rates and allele frequencies (ε_1, ε_2 and ω) to reduce the computational burden of updating these parameters in the Markov chain. When these parameters are known with a high degree of certainty, as may be the case with allele frequencies, then the parentage analysis may be relatively insensitive to this formulation (Nielsen *et al*. 2001). In addition, previous work has also suggested that conclusions

should be relatively robust to sampling variance in the error rates (Wang 2004). In those cases where a high degree of uncertainty is associated with these parameters, it may be prudent to use a range of parameters and check for consistency.

Acknowledgements

We would like to thank Lyanne Brouwer, Nathan Haigh, Paul Johnson, Jessica Metcalf, Shinichi Nakagawa, Richard Nichols, Jeremy Oakley, Mark Rees, Jon Slate and Marco van der Velde. Nature Seychelles kindly allowed us to work on Cousin Island and the Department of Environment and the Seychelles Bureau of Standards gave permission for fieldwork and sampling. This work was supported by the Natural Environment Research Council (NERC), and D.S.R. was supported by an NERC fellowship.

References

Adams WT, Griffin AR, Moran GF (1992) Using paternity analysis to measure effective pollen dispersal in plant populations. *American Naturalist*, **140**, 762–780.

Almudevar A (2003) A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, **63**, 63–75.

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

Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Burczyk J, Adams WT, Moran GF, Griffin AR (2002) Complex patterns of mating revealed in a eucalyptus regnans seed orchard using allozyme markers and the neighbourhood model. *Molecular Ecology*, **11**, 2379–2391.

Burczyk J, Adams WT, Shimizu JY (1996) Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus attenuata Lemmon*) stand. *Heredity*, **77**, 251–260.

Burland TM, Barratt EM, Nichols RA, Racey PA (2001) Mating patterns, relatedness and the basis of natal philopatry in the brown long-eared bat, *Plecotus auritus*. *Molecular Ecology*, **10**, 1309–1321.

Devlin B, Roeder K, Ellstrand NC (1988) Fractional paternity assignment — theoretical development and comparison to other methods. *Theoretical and Applied Genetics*, **76**, 369–380.

Emery AM, Wilson IJ, Craig S, Boyle PR, Noble LR (2001) Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Molecular Ecology*, **10**, 1265–1278.

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Molecular Ecology*, **6**, 861–868.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall, London.

Griffith SC, Owens IPF, Thuman KA (2002) Extra pair paternity in birds: a review of interspecific variation and adaptive function. *Molecular Ecology*, **11**, 2195–2212.

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

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

Jones B, Clark AG (2003) Bayesian sperm competition estimates. *Genetics*, **163**, 1193–1199.

MacCluer JW, Vandeberg JL, Read B, Ryder OA (1986) Pedigree analysis by computer simulation. *Zoo Biology*, **5**, 147–160.

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

Martin AD, Quinn KM, Pemstein D (2004) *Scythe Statistical Library 1.0*. http://scythe.wustl.edu/wiki/index.php/Main_Page

McCullagh P, Nelder JA (1989) *Generalised Linear Models*, 2nd edn. Chapman & Hall, Cambridge.

Morgan MT (1998) Properties of maximum likelihood male fertility estimation in plant populations. *Genetics*, **149**, 1099–1103.

Morgan MT, Conner JK (2001) Using genetic markers to directly estimate male selection gradients. *Evolution*, **55**, 272–281.

Neff BD, Repka J, Gross MR (2001) A Bayesian framework for parentage analysis: the value of genetic and other biological data. *Theoretical Population Biology*, **59**, 315–331.

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

Pemberton JM, Slate J, Bancroft DR, Barrett JA (1995) Nonamplifying alleles at microsatellite loci — a caution for parentage and population studies. *Molecular Ecology*, **4**, 249–252.

R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. http://www.r-project.org

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

Richardson DS, Jury FL, Dawson DA, Salgueiro P, Komdeur J, Burke T (2000) Fifty Seychelles warbler (*Acrocephalus sechellensis*) microsatellite loci polymorphic in Sylviidae species and their cross-species amplification in other passerine birds. *Molecular Ecology*, **9**, 2226–2231.

Roeder K, Devlin B, Lindsay BG (1989) Application of maximum-likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, **45**, 363–379.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380–390.

Sheehan NA (2000) On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review*, **68**, 83–110.

Slate J, Marshall T, Pemberton J (2000) A retrospective assessment of the accuracy of the paternity inference program CERVUS. *Molecular Ecology*, **9**, 801–808.

Smouse PE, Meagher TR (1994) Genetic-analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (Liliaceae). *Genetics*, **136**, 313–322.

Smouse PE, Meagher TR, Kobak CJ (1999) Parentage analysis in *Chamaelirium luteum* (L.) Gray (Liliaceae): why do some males have higher reproductive contributions? *Journal of Evolutionary Biology*, **12**, 1069–1077.

Sorensen D, Gianola D (2002) *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Thomas SC (2005) The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1457–1467.

Thompson EA (1976a) Inference of genealogical structure. *Social Science Information Sur Les Sciences Sociales*, **15**, 477–526.

Thompson EA (1976b) Paradox of genealogical inference. *Advances in Applied Probability*, **8**, 648–650.

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

Jarrod Hadfield is a postdoctoral researcher working as part of a consortium researching evolutionary, behavioural and conservation issues using the model system provided by the Seychelles warbler. David S Richardson's research focuses on the use of neutral and coding molecular markers to resolve evolutionary questions such as the role of the MHC in mate choice and parasite resistance. Terry Burke heads a research group at Sheffield that uses molecular approaches to address ecological questions, especially sexual selection and life history studies in birds.

Appendix

Estimating allelic dropout and stochastic genotyping error rates

Here we present a Bayesian approach for estimating allelic dropout and stochastic genotyping error rates. We define the error rates for both these processes as ε_1 and ε_2, respectively, in accordance with Wang (2004). ε_1 is the probability that in any heterozygote a single allele will drop out. ε_2 is the probability that an allele will be misscored irrespective of whether the individual is heterozygous or homozygous. We assume that misscored alleles are scored with equal probability as any other allele already recorded in the population at that locus.

Our aim is to estimate the error rates ε_1 and ε_2 from the observed set of genotypes $\mathbf{G}_{obs}$. Information is available for estimating these genotyping error rates when there are multiple observed genotypes per individual and/or when observed genotypes are available for related individuals (Hoffman & Amos 2005). We describe the estimation procedure when data are available on tissue samples that have been genotyped more than once, and those samples are taken from individuals that are assumed to be unrelated. When the samples are taken from related individuals (e.g. mother–offspring) the method is easily extended to incorporate this additional information. We assume that error rates are independent across loci and across individuals, although this assumption could be relaxed if necessary. In particular, we note that the presence of null alleles caused by mutations in the flanking regions of microsatellites will cause nonindependence between errors (Pemberton *et al*. 1995). However, errors generated by this process are invisible if pedigree data are unavailable, as affected heterozygous individuals will consistently be scored as homozygotes.

To make the estimation procedure more tractable we augment the observed data with the true unobserved genotypes (**G**) and the allele frequencies in the population (**w**) (Tanner & Wong 1987):

$$\Pr(\varepsilon_1, \varepsilon_2, \mathbf{G}, \mathbf{w} \mid \mathbf{G}_{obs}) \qquad \text{(eqn 13)}$$

The probability of observing a particular genotype given the true underlying genotype is given by Wang (2004), equations 1 and 2:

$$\Pr(\mathbf{G}^{(obs)}_{\mathbf{u},\mathbf{v}} \mid \mathbf{G}_{w,x}, \varepsilon_1, \varepsilon_2) = \begin{cases} (1-\varepsilon_2)^2 + e_2^2 - 2e_1(1-\varepsilon_2-e_2)^2 \\ (u=w, v=x) \\ e_2(1-\varepsilon_2) + e_1(1-\varepsilon_2-e_2)^2 \\ (u=v=w) \text{ or } (u=v=x) \\ e_2^2(2-\delta_{u,v}) \\ (u \neq w, u \neq x, v \neq w, v \neq x) \\ e_2(1-\varepsilon_2+e_2) \\ (\text{otherwise}) \end{cases} \qquad \text{(eqn 14)}$$

if $w \neq x$ (i.e. the true genotype is heterozygous), and

$$\Pr(\mathbf{G}^{(obs)}_{\mathbf{u},\mathbf{v}} \mid \mathbf{G}_{w,x}, \varepsilon_1, \varepsilon_2) = \begin{cases} (1-\varepsilon_2)^2 \\ (u=v=w) \\ 2e_2(1-\varepsilon_2) \\ (u=w, v \neq w) \text{ or } (v=w, u \neq w) \\ e_2^2(2-\delta_{u,v}) \\ (u \neq w, v \neq w) \end{cases} \qquad \text{(eqn 15)}$$

if $w = x$ (i.e. the true genotype is homozygous). u and v are the nonordered alleles scored, and w and x the true nonordered and unobserved alleles. e_1 and e_2 are defined as $\varepsilon_1/(1+\varepsilon_1)$ and $\varepsilon_2/(k_l - 1)$, respectively, where k_l is the number of alleles recorded at locus l. $\delta_{u,v}$ is an indicator variable taking on the value 1 when $u = v$ and 0 when $u \neq v$, respectively.

We sample the true genotypes (**G**) proportional to the full conditional distribution:

$$\Pr(\mathbf{G} \mid \mathbf{G}_{obs}, \varepsilon_1, \varepsilon_2, \mathbf{w}) \qquad \text{(eqn 16)}$$

using Gibbs updates. The distribution of individual i's genotype at locus l is assumed to be multinomial with a single trial and $k_l(k_l+1)/2$ categories representing the possible nonordered genotypes. The likelihood of the true genotype belonging to any of one of these categories is the likelihood derived from equations 14 and 15 multiplied by the probability of that genotype given the allele frequencies (**w**) and the assumption of Hardy–Weinberg equilibrium.

We implicitly assume that all genotypes (**G**) are conditionally independent given the other parameters in the model. When sampled individuals are related, this is no longer the case and the pedigree structure needs to enter into the likelihood given in equation 16. This can be achieved by evaluating the probability of an individual's genotype conditional on the genotypes of the individual's offspring, parents and spouses (Sheehan 2000; this manuscript).

Likewise, allele frequencies are updated using Gibbs sampling, with the frequencies at locus l sampled from a Dirichlet distribution with k_l categories; the parameters of the Dirichlet distribution are estimated from the true genotypes (see equation 4 in Emery *et al*. 2001). The observed genotypes and error rates do not enter into the likelihood as **w** is independent of $\mathbf{G}^{(obs)}$, ε_1 and ε_2 conditional on **G**. When the sample contains related individuals the allele frequencies need to be updated by considering the genotypes of the base (parental) generation only.

The error rates are updated using a Metropolis–Hastings scheme, where candidate values for ε_1 and ε_2 are chosen from a random uniform distribution on the interval 0–0.3. Candidate values are accepted if the likelihood of the new value is greater than for the old value. If the likelihood is less, the new value is accepted with a probability equal to the ratio of the new likelihood to the old.

The likelihoods are calculated according to equations 14 and 15

$$\Pr(\varepsilon_1, \varepsilon_1 \mid \mathbf{G}_{\text{obs}}, \mathbf{G}, \mathbf{w}) \propto \Pr(\varepsilon_1)\Pr(\varepsilon_2) \prod_{i=1}^{i=n_i} \prod_{l=1}^{l=n_l} \prod_{r=1}^{r=n_{r_{i,l}}} \Pr(\mathbf{G}_{\text{obs}i,l,r} \mid \mathbf{G}_{i,l}, \varepsilon_1, \varepsilon_2), \quad \text{(eqn 17)}$$

with the term **w** not entering into the likelihood, as error rates and allele frequencies are assumed to be independent, conditional on true genotypes. The prior distribution for the error rates, $\Pr(\varepsilon_1)$ and $\Pr(\varepsilon_2)$, are uniform on the interval 0–0.3. The subscript $r_{l,i}$, indicates the rth sample at the lth locus for individual i, all errors being treated as independently distributed. n_i, n_l and $n_{r_{i,l}}$ are the number of individuals in the sample, the number of loci sampled, and the number of times each sample has been genotyped, respectively.

Code is available from the Correspondence and is written in C++ and R (R Development Core Team 2004).

Box 1

Probability theory

Using quite simple probability theory we can see how categorical, fractional and full-probability models differ in their assumptions. To do this, however, we need to understand four basic concepts regarding probability distributions of random variables. For simplicity, we will illustrate these concepts using two variables: relative tail length, denoted y, and relative fecundity, denoted ω. We will also ignore the fact that relative fecundity will have a discrete distribution, and assume that the two traits follow a bivariate normal distribution.

The Joint distribution: $p(\omega, y)$

The joint probability distribution of ω and y describes the complete distribution of both variables. The central column of Fig. 6 represents the joint distribution of y and ω when the two variables are independent (top) and when the two variables are not independent (bottom). The contours represent values of y and ω that

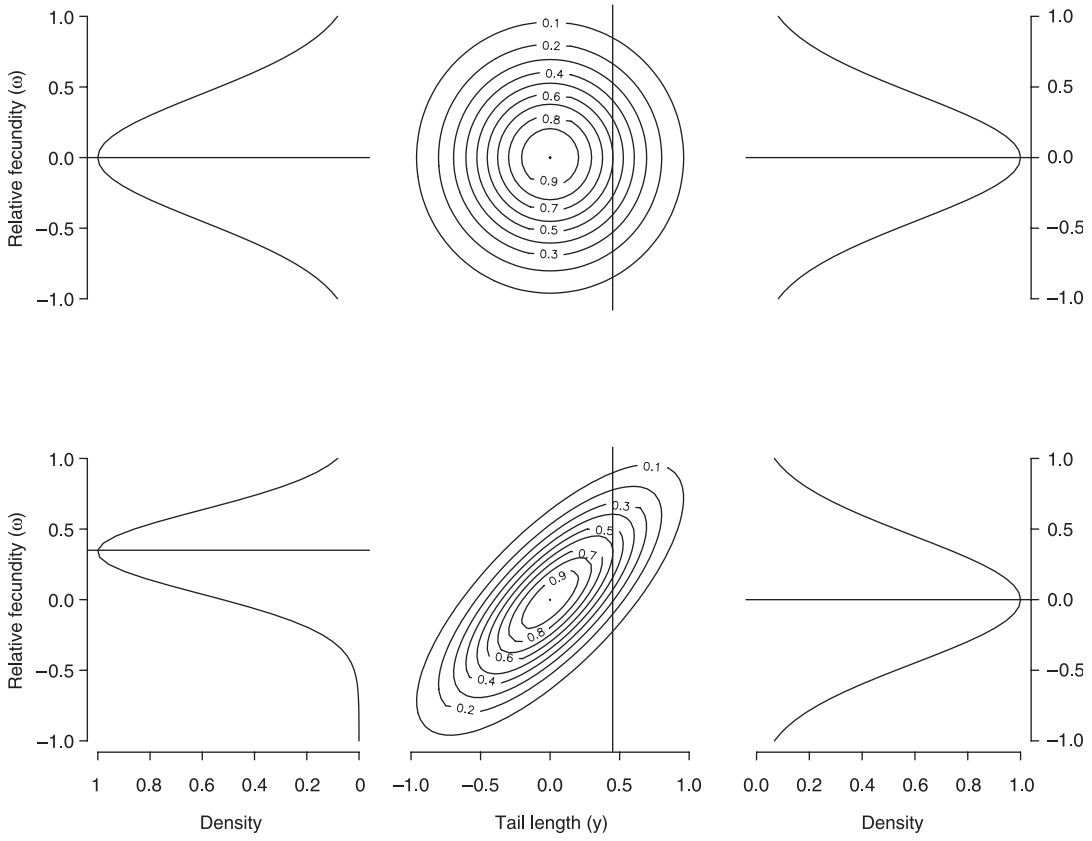

Fig. 6 Distribution of relative fecundity (ωw) and tail length (y) when they are independently distributed (upper plots), and when they are non-independent (lower plots). The central column represents the joint distribution of the two variables. The left hand column represents the distribution of ωw conditional on y = 0.45. The right hand column represents the marginal distribution of ωw.

are equally probable (have the same density), and the 'top of the hill' represents the most likely values ($\omega = 0$, $y = 0$).

The Conditional distribution: $p(\omega \mid y)$
The conditional probability distribution of ω given y describes the distribution of ω for a fixed value of y. The left column of Fig. 6 represents the conditional distribution of ω when $y = 0.45$. This distribution can be visualized as a slice through the joint distribution, indicated by the vertical line. In the upper plot ω and y are independent of each other and therefore knowledge about y gives us no information about ω; knowing the tail length of a male would not tell us anything about how many offspring he has sired. In this example, the conditional distribution of ω given any value of y would be the same. In the lower example, however, ω and y are positively related, so the most likely value of ω when $y = 0.45$ is greater than the most likely value of ω in the joint distribution. Knowing that a male has a tail length of 0.45 reduces our uncertainty in how many offspring he may have sired, and also tells us that he is likely to have sired more offspring than the average.

The Marginal distribution: $p(\omega) = \int p(\omega)dy$
The marginal distribution of ω is the distribution of ω ignoring information about y. The right column of Fig. 6 represents the marginal distribution of ω; it can be visualized, in this instance, as the outline of the joint density 'hill' as seen from the vertical axis. In both examples, we have defined ω and y to have means of zero and variances of 0.2, and consequently both ω and y have the same marginal distributions despite different joint distributions.

Independence
In the above examples, the special properties of independent distributions are immediately obvious. Because either variable provides no information about the other, the marginal distribution of each variable is equivalent to the conditional distribution, $p(\omega \mid y) = p(\omega)$, as can be seen in Fig. 6. For the same reason, the joint density of the two variables is simply the product of their marginal densities $p(\omega, y) = p(\omega)p(y)$. However, when the two variables are not independent the joint distribution is no longer the product of two marginal distributions but is the marginal distribution of one, times the information 'left over' in the other: $p(\omega, y) = p(\omega)p(y \mid \omega) = p(y)p(\omega \mid y)$.

Bayes Theorem
From these concepts we can define Bayes theorem:

$$p(\omega \mid y) = \frac{p(y, \omega)}{p(y)} = \frac{p(y \mid \omega)p(\omega)}{p(y)} \qquad \text{(eqn 18)}$$

Box 2

Probability theory and parentage analysis
We can apply the concepts introduced in Box 1 to see how categorical, fractional and full-probability models differ.

In most cases we are interested in estimating a population-level parameter(s) θ from the genetic **G** and nongenetic **y** data. We are interested in the distribution of these parameters conditional on the data we have collected:

$$p(\theta \mid \mathbf{G}, \mathbf{y}) \qquad \text{(eqn 19)}$$

Typically, θ will concern patterns relating to parentage, such as the relationship between tail length and fecundity in the example given above. Estimating θ directly from the data is difficult. However, if we knew parentage, **P**, without having to rely on genetic data the problem would be simplified, as we could estimate our parameters using well known statistical procedures such as regression and GLMs. To simplify the problem we can therefore estimate the joint distribution of θ and **P** conditional on our data, and then average over the uncertainty in **P**. This allows us to estimate what we want more easily:

$$p(\theta \mid \mathbf{G}, \mathbf{y}) = \int p(\theta, \mathbf{P} \mid \mathbf{G}, \mathbf{y})d\mathbf{P} = \frac{p(\theta, \mathbf{P} \mid \mathbf{G}, \mathbf{y})}{p(\mathbf{P} \mid \theta, \mathbf{G}, \mathbf{y})} \qquad \text{(eqn 20)}$$

The RHS of equation 20 defines our problem completely, but it is perhaps easier to understand when it is framed in terms of Baye's rule:

$$\frac{p(\theta, \mathbf{P} \mid \mathbf{G}, \mathbf{y})}{p(\mathbf{P} \mid \theta, \mathbf{G}, \mathbf{y})} \propto \frac{p(\mathbf{G}, \mathbf{y} \mid \mathbf{P}, \theta)}{p(\mathbf{P} \mid \theta, \mathbf{G}, \mathbf{y})} p(\theta, \mathbf{P}) \qquad \text{(eqn 21)}$$

The numerator on the RHS of equation 21 is the likelihood of us observing the genetic and nongenetic data given **P** and θ, and the denominator takes into account our uncertainty regarding **P**. In a Bayesian setting this is also multiplied by our prior knowledge of **P** and θ before we have observed the data. However, the differences we wish to highlight between the approaches is not the difference between Bayesian and Frequentist

approaches per se, so we will ignore the issue of prior information temporarily.

The fundamental difference between the approaches is the way they in which they choose to define the denominator of equation 21, and hence estimate parentage. From left to right we have the full-probability, fractional and categorical definitions, respectively:

$$p(\mathbf{P}|\theta, \mathbf{G}, \mathrm{y}) \quad p(\mathbf{P}|\mathbf{G}) \quad p(\mathbf{P}|\mathbf{G}) = c \qquad \text{(eqn 22)}$$

where c is a single pedigree configuration. It can be seen from this that although the full probability model defines our problem exactly, the other approaches fail to do so. However, when certain assumptions are met these definitions may be functionally equivalent. Moreover, these assumptions are nested, so that when the categorical definition is true, then so are the definitions proposed by fractional and full-probability methods. Likewise, if the fractional definition is true then so is the full-probability model. The categorical and fractional allocation models are special cases of full probability models.

The assumption made by categorical approaches is that the genetic data are so informative that the probability mass of **P** become centred on a single pedigree configuration, the true pedigree. It is then not necessary to take into account uncertainty in **P** because none exists.

The assumption made by fractional allocation methods is much more subtle. Fractional allocation does allow that the genetic data leave uncertainty regarding **P**, but it does not allow for the fact that the nongenetic data combined with our knowledge of θ can reduce this uncertainty. This assumption is only valid when **P** and **y** are independently distributed conditional on θ:

$$p(\mathbf{P}, \mathbf{y}|\theta) = p(\mathbf{P}|\theta)p(\mathbf{y}|\theta) \qquad \text{(eqn 23)}$$

When the very reason for estimating θ is to see if a relationship exists between the nongenetic data and **P** we contradict ourselves; we maintain that no relationship exists, and then proceed to estimate the relationship. This reduces the power of our tests, because knowledge of θ and **y** can reduce our uncertainty in **P**, and also adds bias, because the reduction in uncertainty in **P** does not translate into a reduction in uncertainty in θ that is symmetrical.

To take the example involving fecundity and tail length: when $\theta = 0$, there is no relationship between fecundity (a property of **P**) and tail length, and the upper plots of Fig. 6 illustrate the situation. There would be little advantage in using the nongenetic data (tail lengths) to help us infer the distribution of fecundities in the population, because they are not related. The marginal and conditional distributions of ω are identical. However, if $\theta \neq 0$ then a relationship does exist between fecundity and tail length, and then the lower plots in Fig. 6 illustrate the situation. If we ignore the tail lengths when we want to infer the fecundity of individuals, we would be using the marginal distribution of ω. This may be very different from the conditional distribution, which is the distribution we should be working with, since we do know the tail lengths of the sampled males. The conditional distribution of ω may have a different mean than the marginal distribution of ω, and generally has greater precision.

In conclusion, full-probability models are an exact solution to the problem at hand. However, if the genetic data leave no uncertainty about parentage then categorical and fractional approaches converge on this exact solution. When there is uncertainty, the fractional approach converges on this exact solution, but only under certain restrictive conditions. These conditions preclude us from entertaining notions of heritability and nonrandom mating, the very basis for evolution.