Genetics: Published Articles Ahead of Print, published on June 11, 2007 as 10.1534/genetics.106.067066

Estimating Differential Reproductive Success from Nests of Related Individuals, with Application to a Study of the Mottled Sculpin, *Cottus bairdi*.

Beatrix Jones[*], Gary D. Grossman[†], Daniel C. I. Walsh[*], Brady A. Porter[‡], John C. Avise[§], Anthony C. Fiumera[**,††]

[*] Institute of Information and Mathematical Sciences, Massey University-Albany, Albany, New Zealand

[†] D. B. Warnell School of Forestry & Natural Resources, University of Georgia, Athens, GA, 30602, USA

[‡] Department of Biological Sciences, Duquesne University, Pittsburgh, PA, 15282, USA

[§] Department of Ecology and Evolutionary Biology, University of California-Irvine, Irvine, CA 92697, USA

[**] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

[††] Current address: Department of Biological Sciences, Binghamton University, Binghamton, NY 13902, USA

Running header: Parentage Estimation for Nest-Structured Data

Key Phrases: Parentage analysis, Reproductive success, Cluster sampling,, Age structure, Bayesian Estimation

Corresponding Author: Beatrix Jones

Institute of Information and Mathematical Sciences

Massey University-Albany Campus

Private Bag 102-904

North Shore Mail Centre

Auckland, New Zealand

Email: m.b.jones@massey.ac.nz

Telephone: +64 9 414 0800, ext. 41064

Fax: +64 9 441-8136

ABSTRACT

Understanding how variation in reproductive success is related to demography is a critical component in understanding the life history of an organism. Parentage analysis using molecular markers can be used to estimate the reproductive success of different groups of individuals in natural populations. Previous models have been developed for cases where offspring are random samples from the population but these models do not account for the presence of full- and half-sibs commonly found in large clutches of many organisms. Here we develop a model for comparing reproductive success among different groups of individuals that explicitly incorporates within-nest relatedness. Inference for the parameters of the model is done in a Bayesian framework, where we sample from the joint posterior of parental assignments and fertility parameters. We use computer simulations to determine how well our model recovers known parameters and investigate how various data collection scenarios (varying the number of nests or the number of offspring) affects the estimates. We then apply our model to compare reproductive success among different age groups of mottled sculpin, *Cottus bairdi*, from a natural population. We demonstrate that older adults are more likely to contribute to a nest, and that females in the older age groups contribute more eggs to a nest than younger individuals.

INTRODUCTION

Parentage analyses via molecular markers can be used to investigate a variety of demographic, behavioral, and evolutionary parameters in natural populations (e.g., Avise *et al.* 2002). For example, researchers have used genetic markers to determine the rate of extrapair fertilizations in 'socially monogamous' species (Birkhead and Møller 1995) and estimate the number of fathers contributing to the clutch of a single female (Myers and Zamudio 2004). Parentage studies also can be used to estimate reproductive success among potential parents (Smouse and Meagher 1994), and parameters such as the effective number of breeders within a population (Fiumera *et al.* 2002), or they can be combined with quantitative genetic analyses to identify quantitative trait loci (QTL) in natural populations (Slate *et al.* 2002) and to estimate heritabilities through analysis of wild caught females and their progeny (King *et al.* 2001). Parentage studies can also be applied to address questions relating to gene flow and dispersal (Burczyk *et al.* 2006).

Parentage studies also show promise for estimating differential reproductive success among individuals within populations. Nielsen *et al.* (2001) used this approach to compare the reproductive success of dominant and subordinate males in North Atlantic humpback whales. A multitude of questions in evolutionary and conservation biology can be addressed with this approach: examples include estimating whether there are differences in reproductive success between nest tending or cuckolding males (Neff *et al.* 2000), resident versus immigrant males (Johannesen and Andreassen 1998), freshwater versus anadromous trout (Curry 2005), wild versus hatchery reared fish (Dannewitz *et al.* 2004), or different age classes of individuals that contribute to particular nests (Røed *et al.* 2005).

In parentage analysis, genotypic information is collected from offspring and their potential parents. The goal may be determination of the true mother and father as in CERVUS (Marshal *et al*. 1998) or FAMOZ (Gerber *et al.* 2003), with post-hoc inference for demographic parameters. Alternately, conclusions can be drawn about the parameters of interest using a model likelihood or posterior that incorporates all possible parental assignments (e.g. Roeder *et al*. 1989; Adams *et al.* 1992; Nielsen *et al.* 2001; Jones 2003). Both approaches have been developed assuming that progeny are a random sample from the population. Although this assumption may be reasonable for species that produce only a single offspring within a reproductive bout (e.g., Nielsen *et al.* 2001), or for broadcast spawners where offspring may mix randomly (e.g., Levitan 2005), in many other species related progeny are clustered into groups that are more likely to be full- or half-sibs than offspring randomly selected from the population. Examples include fish nests that typically are guarded by the male parent (DeWoody *et al.* 2000b), or litters of pups (Shurtliffe *et al.* 2005) or egg strings (Emery *et al.* 2001; Walker *et al.* 2007) that are produced by a single female. The availability of related siblings offers unique opportunities for parentage analysis, but current statistical methods are not well suited to analyzing brood-structured data.

Sieberts *et al.* (2002) and Nason *et al*. (1998) have shown that siblings considered jointly contain much more parentage information than offspring considered singly. The presence of multiple progeny from a single parent may allow the full multilocus genotypes of the parents to be determined (making parental assignments more reliable), but the design and analysis of studies using groups of related progeny are more complex than the random sampling case. In particular, if the relatedness of offspring within a nest is ignored, the variance of reproductive success estimators across groups of parents will be underestimated. Despite the potential for

using progeny arrays for parentage assignment, the techniques currently available are generally confined to partitioning a set of offspring into full- and half-sibships (Butler *et al.* 2004). These techniques also do not consider the genotypes of putative parents (although in some cases such as COLONY [Wang 2004] they can reconstruct the likely parental genotypes). An exception is PARENTAGE (Emery 2001), which can use information on potential parents in reconstructing sibships for a single nest of progeny.

Here we develop a model for comparing reproductive success among different groups of individuals that explicitly incorporates within-nest relatedness. Inference for the parameters of the model is done in a Bayesian framework, where we sample from the joint posterior of possible parental assignments and fertility parameters. We then use simulated data to establish the ability of our method to recover known parameters, and we suggest optimal data collection strategies. Finally we apply our approach to compare the reproductive success of different age groups of individuals in a natural population of the mottled sculpin, *Cottus bairdi*, a freshwater fish common to small streams in the eastern United States.

METHODS

We developed a general model for the reproductive success of different categories of individuals. This model is then extended to encompass the genotype probabilities for sampled nests and putative parents. This allows us to use observed genotype data to generate a joint posterior for parent assignments and fertility parameters. A Markov chain Monte Carlo algorithm is used to characterize this posterior via sampling. Our model was developed for a parentage data set derived from a natural population of mottled sculpin, *Cottus bairdi* (Fiumera *et al.* 2002) with the goal of comparing the reproductive success of two different age groups of

males and females. In mottled sculpins, males guard a nest where multiple females may deposit their eggs. Thus a single nest can be composed of half- and full-sibs. A sample of the progeny from multiple nests and any potential parents are genotyped at codominant genetic markers (e.g., microsatellites). The genotypes are then used to assess paternity and maternity among the putative parents. We then used simulated data sets, where the true model parameters are known, to assess the accuracy and precision of our approach under conditions consistent with the mottled sculpin data set. Finally, we applied our Bayesian estimator to the sculpin parentage data set and compared our estimates of variation in reproductive success to those obtained using the parentage assignments from Fiumera *et al.* (2002) as well as the programs COLONY (Wang 2004) and PARENTAGE (Emery *et al*. 2001). We term the estimates from Fiumera *et al.* (2002) the 'BY EYE' estimates, as the parentage assignments were made via investigator inspection of the genotypes.

Model of Reproductive Success

Our Bayesian approach estimates seven population level parameters using the offspring and putative parent genotypes (Table 1). The number of offspring produced by parents in an age class *i* is affected by: 1) the probability that a spawning parent is from age class *i*, and 2) the fraction of offspring typically spawned in a nest by a particular parent from age class *i* (when there are multiple parents of the same sex). In our model, a mother participating in a nest is from group *i* with probability α_{iM}, with the α_{iM} constrained to sum to 1; α_{iF} are the analogous parameters for fathers. The total number of mothers participating in a nest is a truncated Poisson with parameter λ (where the truncation removes the possibility of zero mothers) and the number of fathers is geometric with parameter p. If p is close to one, most nests have exactly 1 father. The parameter γ_i governs the fraction of offspring produced by mothers in age class *i*; the γ_i are

constrained to sum to 1, and equal γ_i represent equal production across age classes. The number of offspring belonging to a particular father depends on the father's cuckolding status rather than age class. When multiple males contributed to a nest, the one with fewer offspring was considered the cuckolder. Offspring belong to a cuckolding father with probability β/k, where k is the number of cuckolding fathers assigned to that nest. The focus on different age classes is particular to the mottled sculpin application, but is relevant to other iteroparous organisms with multi-year lifespans. The different categories could be any designation appropriate to the species or question of interest and by constraining some parameter values this model can easily be adapted to other mating systems, including those where one or both sexes are monogamous. A detailed description of the model follows.

Imagine a nest with n_O offspring with genotypes O_h, n_M total mothers (n_{Mi} in each age classes, with n_F, n_{Fl} similarly defined for the fathers); genotype M_{ij} for the ith mother in age class j, primary father genotype F_1 and other father genotypes F_k. The probability an offspring belongs to a particular mother in age class i is:

(1) $\gamma_i / \sum_j n_{Mj}\gamma_j$.

with the γ_i constrained to sum to one. The probability for the entire nest is then:

(2)

$$\prod_{\mathrm{i}} \alpha_{Mi}{}^{n_{Mi}} \times \prod_{\mathrm{l}} \alpha_{Fl}{}^{n_{Fl}} \times \mathrm{Geom}(n_F \mid p) \times \mathrm{tPois}(n_M \mid \lambda) \times$$

$$\prod_{h=1}^{n_O} \left[(1-\beta)^{1(n_F>1)} \sum_i \left\{ \frac{\gamma_i}{\sum_q n_{Mq}\gamma_q} \sum_{j=1}^{n_{Mi}} P(O_h \mid M_{ij}, F_1) \right\} + \frac{\beta}{n_F - 1} \sum_{k=2}^{n_F} \sum_i \left\{ \frac{\gamma_i}{\sum_q n_{Mq}\gamma_q} \sum_{j=1}^{n_{Mi}} P(O_h \mid M_{ij}, F_k) \right\} \right],$$

where the second term in the sum appears only when there are cuckolding fathers. Geom(.|p), and tPois(.|λ) denote the geometric and truncated Poisson, and geometric probability mass functions with parameters p and λ respectively and $P(O_h|M_{ij}, F_k)$ is the segregation probability.

We now imagine that we can also observe which of the nest parents are among our captured individuals. Let I_{Mi} be a vector of indicator variables with length n_{Mi}; I_{Mjj}, the jth entry for this vector, is m when the jth mother from age class i corresponds to the mth captured mother; otherwise it is zero. An analogous vector exists for males. Each captured adult can appear only once. Let $|I_M|$ and $|I_F|$ be the number of captured mothers and fathers (i.e., the number of non-zero entries in the I_M and I_F vectors), $ij \in \{I_{Mij}=0\}$ and $k \in \{I_{Fk}=0\}$ denote the indices of uncaptured parents, and $f(M_{ij})$, $f(F_k)$ the population frequencies of genotypes M_{ij} and F_k. The probability for the fully observed nest (including the genotypes for uncaptured parents) and the capture vectors is then the expression in (2) times

$$(1-g_0)^{|I_M|+|I_F|} g_0^{n_N+n_F-|I_M|-|I_F|} \prod_{ij\in\{I_{Mij}=0\}} f(M_{ij}) \prod_{k\in\{I_{Fk}=0\}} f(F_k).$$

The likelihood for many nests is taken to be the product of the individual nest likelihoods, (i.e., the nests are independent). Thus there is no constraint that a captured parent can appear in only one nest. However, our likelihood essentially reflects a separate "capture" factor of 1-g_0 each time an individual appears. This deviation from reality will be minor if participation in multiple nests is rare; a more sophisticated model would be necessary in other cases.

In practice, the n_M, n_{Mi}, n_F, n_{Fl}, M_{ij}, F_k, I_M, and I_F are unknown; these are treated as nuisance parameters over which we must integrate. The configurations of these variables with non-zero likelihood is constrained by the observed data; for instance, if I_{Mij} is 33, M_{ij} must match the observed genotype for the 33rd captured mother. The indicator vectors I_M and I_F constrain but do not fully determine n_{Mi} and n_{Fl}, as only the count in each age class among the *captured*

parents can be computed from the indicator vectors. Rather than using additional latent variables to represent the age classes of uncaptured individuals, we use the following simplification: the uncaptured group is assigned parameter $\gamma_0 = \sum \alpha_{iM} \gamma_i$, and this value is used in (1) to determine the probability an offspring comes from one of the unobserved mothers assigned to its nest. We make a corresponding modification to the multinomial distributions in (2), which now have g_0 as the probability of an uncaptured parent; the other probabilities are renormalized to sum to $1-g_0$.

The allele frequencies at each locus are assumed to be known; population genotype frequencies $f(M_{ij})$, $f(F_k)$ are computed assuming Hardy-Weinberg and linkage equilibrium. Typing error is another important consideration in practice. Our segregation probability incorporates a simple model of typing error for the offspring; the offspring's genotype has a specified probability of being erroneous at each locus. The erroneous genotype is drawn at random from the population frequencies. We have not modeled typing error in the parents. If a true parent were mistyped it would result in an 'uncaptured' parent being assigned in the inferred family, resulting in a loss of valuable data. However, it should be very unlikely that an erroneous individual (even if mistyped at a single locus) is considered a true parent if the genetic markers used have reasonable exclusion power.

In our Bayesian treatment, the parameter vectors α_{iM}, α_{iF} and γ_i have uniform dirichlet priors. The other parameters have uniform priors tailored to the mottled sculpin example: the truncated Poisson parameter λ is uniform on (0, 10); p on (0.5, 0.98); β on (0, 0.5), incorporating the assumption that the primary father will have the majority of the offspring; and g_0 is uniform on (0,1). These are all easily changed to fit other situations.

We fit the model using a Metropolis-Hastings algorithm (Hastings 1970) which samples from the joint posterior of these unknowns and the parameter values. Details are given in the

appendix; the program source code is available from http://www.massey.ac.nz/~mbjones/research/. Runs of 2.5 million iterations, in which the parameters were sampled every 500 iterations, were found to be adequate. Under these conditions, the Monte Carlo standard error (the difference in estimates when the algorithm is run with different seeds) is small compared to the posterior standard deviation. Five runs with different seeds using one of the 5 locus populations described below showed that the Monte Carlo standard error was less than 10% of the posterior standard deviation for most parameters. The exceptions were α_M (13% of the posterior standard deviation) and g_0 (21% of the posterior standard deviation). These parameters, however, have small posterior standard deviations, so the Monte Carlo errors are still quite small in absolute terms—about 0.01 for parameters that can range between 0 and 1. Autocorrelation (and therefore Monte Carlo variance) properties were found to be similar for all chains run, despite the differences in the posterior distributions from which they were sampled.

Mottled Sculpin Data

During the breeding season, male mottled sculpin defend nest rocks where females deposit the eggs and the males guard the eggs until hatching (Savage 1963). Fiumera *et al.* (2002) genotyped 1,259 offspring from 23 nests and 455 juveniles and adults at 5 microsatellite DNA markers. The number of alleles (and observed heterozygosity) for the loci was 4 (0.58), 8 (0.74), 9 (0.81), 16 (0.64), and 23 (0.85). At least 48 offspring (or all the offspring in the two cases where fewer than 48 existed) were genotyped from each nest and one nest was exhaustively sampled, with 209 of the 210 offspring successfully genotyped. Fiumera *et al.*

(2002) estimated (using a 'genetic' mark-recapture approach) that between 47% and 75% of the putative parents were collected. Of the captured adults, 43% were male.

The 455 juveniles and adults were aged using the methods of Grossman *et al.* (2002). In brief, after clearing saggital otoliths with cedar wood oil we identified annual bands using a dissecting scope and reflected light**.** Female mottled sculpin are moderately long-lived with a maximum recorded lifespan of 7+ years in the Coweeta Creek drainage (Grossman *et al.* 2002; Figure 1). Thus there are many reproductively active age classes.

Fitting separate parameters for each age class is not feasible with the amount of data available. To reduce the number of parameters estimated, age classes were binned into two age groups (see below). The selection of which age classes to bin has consequences for the interpretation of the model and the power to detect differential reproductive success. The α parameters for both males and females now represent the probability of nest participation aggregated over age classes in the same group. Differences between α and the frequency of a group in the population indicates differential nest participation between groups; however poor choice of groups (e.g., grouping the most likely to reproduce age class with the least likely) could obscure these differences. The γ parameters are also age class dependent. Under grouping of age classes, the average fraction of offspring in a nest attributed to a particular age group will be the same as predicted by the model using a γ averaged over nest-participating individuals in the group. The true variance of the offspring fraction will be slightly larger than implied by the model with the averaged γ, but this effect is small for the range of γ relevant in this problem. Again, poor choice of groups could obscure reproductive differences. Estimation of γ also relies on co-occurrence in nests of mothers of different age groups, so each age group must participate in nests often enough to make this a common occurrence.

Reproductive output for both female and male sculpin increases with age. This occurs for males via multiple matings, and for females via increased fecundity (Grossman *et al.* 2002). Thus, it is probably appropriate to group sculpin in adjacent age classes, so we binned age classes 2 and 3 into "group 1" and age classes 4 and older into "group 2" (Figure 1). This method of grouping placed 64% of females in age group 1 and 36% in age group 2. We excluded age class 1 individuals for three reasons: females of this age are a mixture of reproductive and immature individuals (Grossman *et al.* 2002); Fiumera *et al.* (2002) found little evidence for genetic parentage by such young sculpins; and initial runs of our current method likewise indicated little reproductive involvement by these fish (data not shown). Excluding these individuals allows us to focus on characterizing differences between age classes where all individuals are capable of reproduction.

Simulation Study

We use computer simulations to illustrate the ability of our program to recapture true parameter values for the model under a variety of conditions tailored to the mottled sculpin population. The allele frequencies at the five loci used for the simulations were based on the sample of 455 individuals (~350 adults) in Fiumera *et al.* (2002). We simulated the data with an error rate of 0.01, which incorporates both novel mutations occurring between the parent and offspring and also the possibility of genotyping errors. We fix the expected number of observed adults at 350 with males comprising 43% of the population. We model two age groups, with 70% of the population in the younger age group (see Figure 1). Unless otherwise noted, there are 22 sampled nests of 48 eggs each, and the adult population consists of 700 individuals (i.e., 50% of the adult population has been observed). The parameter values used to generate the data

are presented in Table 1. Performance of the algorithm is measured by the bias and variance of the posterior; ideally, the posterior samples will be tightly clustered around the true values.

Each simulated population and subsequent sample of nests and adults was created by randomly sampling from the actual distributions defined by the true parameter values. Thus, within a given replicate population the observed quantity corresponding to a parameter could vary from the true value (e.g., the observed fraction of mothers from age group 1 will not be exactly 0.58, the value used for α_M in the simulations). This is a consequence of sampling only a finite number of nests (22), analyzing only a finite number of progeny (48) and collecting only a subset of the actual adults (i.e., only about half of the deduced parents will provide data for estimation of the age group parameters). One important question is how much each replicate varies from the true parameter because of this limited sampling. We can address this by estimating the parameters for each replicate using the full parentage information that is known from the simulations. We can then compare these parameter estimates to those obtained by fitting our Bayesian model to the observed genotype data, where neither the true parentage nor the parameter values are known. Thus we can gain some information about how uncertainty in parentage inference affects the parameter estimates.

First we investigated how finite sampling affects the variance in parameter estimates. Fifteen populations were generated under the conditions described above and the parameters were estimated using the known parentage from the simulations. We then investigated the performance of our MCMC Bayesian approach to estimate the parameters using the genotype information from the offspring and parents when the true parents are not known. Genetic loci were simulated for both the parents and offspring (either , 4 or 5 loci were simulated, with 5 populations assigned to each condition). In each case, the least polymorphic loci were used to

show the maximum changes in variance. We then ask how well the parameter estimates from our MCMC Bayesian model agree with the parameter estimates that were calculated using the known parentage information.

Next we considered different ways of increasing data (in each case, essentially by 50%). The impact on the uncertainty for each parameter was then measured and compared to the mean standard deviation for a population with 22 nests, 48 offspring per nest, and 50% of the parents typed. Parameter uncertainty was measured by the posterior standard deviation averaged across the simulated populations. The properties of this quantity are well known for simple estimation problems, enabling instructive comparisons. The ultimate quality of our estimates is of course affected by other factors as well, including bias and Monte Carlo error.

All simulations in this set use 5 loci. First, we simulated 5 populations where the proportion of adults that had been genotyped was increased to 75%. This was accomplished by decreasing the total size of the simulated population to 467 individuals, so that the 350 that were genotyped constitute 75% of all adults. We then simulated 5 populations where the number of nests was increased to 33; and 5 more where the offspring per nest was increased to 72. We then considered the possibility of increasing the number of nests analyzed to offset a decrease in the percentage of parents that were genotyped. This was accomplished by simulating 5 populations where 48 offspring from each of 44 nests were analyzed but only 25% of the adults were genotyped.

A third set of simulations considered strategies for additional genotyping within nests that showed multiple maternity, the attempt being to improve estimates of the fraction of nestmates produced by dams of different age groups ($\gamma_\cdot$). For these simulations, we used a single nest and considered the posterior for γ only. We simulated 5 replicates of a single nest with 96 sampled

offspring under each of the following configurations (unless noted, the mothers are assumed to be among the typed adults, and there is a single un-typed father): 2 mothers of different age groups; 4 mothers, one of which has a different age group; 4 mothers, 2 from each age group; and 4 mothers, 1 from each age group and 2 un-typed and un-aged.

Finally, a fourth set of simulations considered null alleles. Although explicit modeling of null alleles is possible (and essential in cases where they are at high frequency), it is not undertaken here. Rather, we assess the robustness of our algorithm to ignored null alleles. We simulated 5 replicate populations where one locus has a null allele with frequency 10%. We then fitted the model to the observed data, ignoring the presence of the null allele, and compared estimates to those based on the complete parentage information.

Application to Mottled Sculpin Data Set

We applied our MCMC Bayesian approach to estimate the seven parameters determining reproductive success (Table 1) for the actual mottled sculpin data set. We then used these parameter estimates to ask whether: a) females from the older age group are more likely to be mothers (i.e., does α_{1M} differ from the frequency of females in age group 1?), b) females from the older age group produce a greater proportion of the offspring in a nest (does γ_1 differ from 0.5?); and c) are males from the older age group more likely to be fathers (does α_{1F} differ from the frequency of males in age group 1?). We assumed an error rate of 0.01 (which incorporates novel mutations occurring between parent and offspring, as well as genotyping errors). Two nests had indications of null alleles at one locus, and the genotypes at that locus were treated as missing data for the affected individuals.

We then compared our results to the 'BY EYE' approach used by Fiumera *et al.* (2002), as well as to results from the programs COLONY (Wang 2004) and PARENTAGE (Emery

2001). Fiumera *et al.* (2002) reconstructed multilocus parental genotypes by inspection and assigned parentage if the full multilocus genotype of an adult matched the reconstructed parental genotype. The authors allowed for novel mutations or genotyping errors based on the investigator's judgment, and the final assignments invoked a conservative error rate of 0.002. As we have used them, COLONY and PARENTAGE inferences are also based on matches between typed individuals and inferred parental genotypes.

COLONY is designed to look only at the offspring, and partition them into full-sib groups nested within half-sib groups; parental genotypes are then reconstructed based on these groupings. To incorporate the information that offspring in different nests are unlikely to be related, we ran COLONY separately on the data from each nest, specifying the population allele frequencies; an error rate of 0.01 was also incorporated. We then looked for matches between the likely parental genotypes inferred by COLONY and the typed parent individuals. A match between any of the multilocus genotypes specified by COLONY leading to the maximum likelihood (up to 32 genotypes) and a typed individual was used; there were no instances where more than one of the COLONY inferred genotypes for an individual matched genotyped adults in different age groups (ambiguity between parents in the same age group occurred in only one case, and did not affect parameter estimates).

PARENTAGE was also run separately for each nest. While PARENTAGE has the ability to consider the genotypes of putative parents in inferring family structure, prior specification was found to be difficult when this option was used. Runs with several different priors were done, with none found to be suitable (results not shown). Instead, the putative parents were disregarded and the priors outlined in Emery *et al.* (2001) were used (with the roles of the sexes reversed to accommodate the mottled sculpin mating structure). Posthoc matching

of captured parents with the inferred parental genotypes for the maximum a posteriori family configuration of each nest was then performed, a process similar to that used for COLONY.

For the COLONY, PARENTAGE and 'BY EYE' methods, we took the inferred parent to offspring assignments as fixed, i.e. we set $P(O|M,F)$ to be zero for all M,F pairs other than the one assigned. Under these conditions, the likelihood in equation (2) factors into separate terms for λ, g_0, α_F, β, and p; and a term involving γ and α_M. Thus, the posterior is a product of independent univariate posterior densities (and one bivariate density). Using the same priors specified for the model, we found posterior densities based on the 'BY EYE,' COLONY and PARENTAGE assignments. This was done analytically in the cases where the priors are conjugate, and by calculating the likelihood on a fine grid and normalizing in other cases.

RESULTS

Sculpin Age Data

Of the 455 post-larval individuals that we genotyped, 426 were both successfully aged and sexed. Immature individuals were not considered in the analysis. In total, 338 individuals were at least two years old and thus potentially reproductive. Among these, we unambiguously sexed 328 individuals; ambiguous individuals were considered both as potential mothers and potential fathers. As previously noted, ages were binned into two age groups. Age group 1 comprised age classes 2 and 3 years, and age group 2 comprised age classes 4 and older (Figure 1). Age group 1 comprised 64% of the female population and 79% of the male population. Population allele frequencies were treated as known, and calculated from all 455 genotyped individuals.

Simulations

We used computer simulations to investigate how precisely we could recover the true parameter values under conditions similar to those from the mottled sculpin data set. Given that we knew the true parentage assignments, we could examine the deviation of the parameter estimates calculated using perfect knowledge of the true parents and measure the performance of our Bayesian approach when we do not know the true parents. Thus, we could gain information regarding how much of the deviation was due to having only sampled 22 nests and 50% of the adult population versus how much of the deviation was due to the parentage inference procedure.

Even when the true parents were known perfectly, some of the parameter estimates had large variances around the true means (boxplots in Figures 2 and 3). This was especially evident for the male mating parameters. Estimates of α_{1F} using the known parentage ranged from 0.33 to 0.77; the simulation value was 0.57. Most nests have only one male parent, so α_{1F} is typically estimated with around 11 fathers that have age information. Estimates of the proportion of offspring sired by a cuckolding male (β) was also affected because it relied on observing a nest with more than one male parent. Given that the probability of cuckoldry was small, some simulated populations did not have any nests of this type, so the inferred value for β in these replicates is the prior mean of 0.25 (closer inspection of the posterior would reveal that it retains a uniform distribution between 0 and 0.5). By contrast, many of the female parameters had much smaller error variances because, on average, almost three times as many females contributed to each nest and, thus, these parameters were typically estimated using many more informative data points. Below we investigate data collection scenarios that could be used to increase the amount of information available for the parameter estimates, but first we assess how well our MCMC Bayesian method agreed with the estimates obtained using the full parentage information.

Overall, our MCMC Bayesian approach performed well at recovering the known parentage estimates when the true parental relationships were unknown (dashed lines, Figures 2 and 3). In general, the deviation of the MCMC Bayesian estimate from the known parentage estimate decreased as the number of loci increased. We note underestimation of g_0 with three loci, because limiting the number of loci resulted in some erroneous matches with observed parents (Figure 3D). These erroneous matches were more likely to be with individuals in the more common younger age group, resulting in a mild upward bias for both α_1 and γ_1. There was also overestimation of g_0 with five loci, as a result of ignoring parental typing errors that resulted in the exclusion of some true parents. Even when our Bayesian approach did well at recovering known parentage, the deviations between our method and the true population values still had a large range (solid lines, Figures 2 and 3). By comparing the solid lines with the box plots in each panel (representing the estimates using complete parentage information) we find that this deviation was largely due to having sampled only a finite number of nests and adults (rather than uncertainty in the parentage assignments).

Given the limitations imposed by the original data, we evaluated how the precision of the parameter estimates might be affected by different data collection scenarios. Improvement in the precision of our Bayesian estimator was measured via the resulting change in the standard deviation of the posterior across the different scenarios. We compare posterior standard deviations by looking at the ratio of the average posterior standard deviation under the altered sampling scheme to the average posterior standard deviation of the original sampling scheme (Figure 4). Increasing the proportion of parents genotyped (to 75%), the number of nests analyzed (to 33), or the number of offspring analyzed per nest (to 72) each effectively represents a 50% increase in the amount of data. In simple situations, this should decrease the standard

deviation by a factor of 1/sqrt(1.5); a dashed line marks this level of improvement in Figure 4. Note, in Figure 4, that the standard deviation of g_0 (the proportion of parents typed) is not reported because the value for that parameter changed across the different data collection scenarios investigated; this was the primary driver for observed changes in the posterior standard deviation of g_0.

Parameter estimates that depend on knowing the age group of assigned parents (α_M, α_F, γ) were improved by increasing the percentage of parents genotyped or increasing the number of nests analyzed (Figure 4A). Estimation of the relative fecundity of the different age groups of females (γ) was also improved by increasing the number of offspring analyzed per nest. Estimates that depend upon knowing the number of parents in the nest (λ, p) were improved only by increasing the number of nests analyzed (Figure 4B). As expected, increasing the number of nests improved the estimates of every parameter, and the estimates of β were improved with every data collection scenario investigated. It is notable that increasing the number of nests analyzed can largely compensate for the reduced proportion of adults that were genotyped (lightest bar in Figure 4). With 44 nests analyzed but only 25% of the adult population genotyped, the standard deviations were, at worst, just above those for the reference scenario.

Next we investigated how additional typing of offspring from select nests with multiple maternity might improve estimates of the relative fecundity of mothers from different age groups (γ.). Here we compare among strategies for selecting nests for additional typing (all would improve on typing only 48 offspring per nest). There was little difference in the posterior standard deviations obtained from increased genotyping for nests with one genotyped mother in each age group (0.048), or for nests with three genotyped mothers in age group 1 and one in age group 2 (0.050). The average posterior standard deviation for the case with two mothers in each

age group was slightly larger (0.067) due to one replicate with a multimodal posterior for relative fecundity ($\gamma_\cdot$). However, when there were two mothers in each age group but two of them were not genotyped or aged, then the posterior standard deviation of γ was larger (mean 0.092 across all replicates).

The effect of null alleles on most of the parameter estimates was small (data not shown) but null alleles did impact estimates of g_0 and p. Ignoring null alleles resulted in underestimating the proportion of adults that actually were genotyped (i.e., g_0 was overestimated), and it resulted in overestimating the frequency of cuckoldry (i.e., p was underestimated). This likely occurs because ignoring null alleles results in some true parents being erroneously excluded.

Application to Sculpin Data Set

We then used the actual sculpin data set from Fiumera *et al.* (2002) to generate the posterior for the seven parameters defining reproductive success using our MCMC Bayesian approach. We examined the following hypotheses. Are individuals in the older age group over-represented among the parents, or, phrased another way, does the proportion of nest-participating individuals from age group 1 differ from the population proportion (i.e., is $\alpha_{1M} < 0.64$ or $\alpha_{1F} < 0.79$)? Also, do age group 1 females produce fewer offspring than older females (i.e., is γ_1 less than 0.50)? The posterior probability $\alpha_{1M} < 0.64$ (estimated from the proportion of sampled α_{1M}'s less than 0.64) proved to bewas 0.95; the posterior probability $\alpha_{1F} < 0.79$ proved to bewas 0.96. These results strongly suggest that males and females in the older age class enjoy increased parental representation in nests. The sampled values for γ_1 also were entirely below 0.5—our estimate for the posterior probability of $\gamma_1 < 0.50$ was 1. These data demonstrate conclusively that older females produce a larger fraction of the offspring for the nests in which they participate.

In general, our MCMC Bayesian approach yielded estimates consistent with those obtained 'BY EYE,' COLONY, and PARENTAGE (Table 2). We restrict comment to cases where two methods differed by more than two standard deviations (using the smaller of the two standard deviations). The largest discrepancy was for the parameter β, which defines the proportion of offspring sired by a cuckolding male. The 'BY EYE' method estimated β to be substantially larger compared to the other approaches. Remember, however, that all approaches estimated the rate of cuckoldry to be very low (i.e., p is close to 1), such that β is estimated using a very limited amount of data. The model fit in each case was also limited in the sense that cuckoldry was the only mechanism modeled that could account for different fathers contributing to the same nest (by whatever method this model is fit, multi-father nests will increase the "cuckoldry parameter" p). Inspection of the mottled sculpin data suggests that a nest takeover is a more plausible explanation for the single nest with large contributions from multiple fathers.

The estimates of α_{1M} from our MCMC method were smaller than those for all other methods (although the difference exceeds two standard deviations only for COLONY). Conclusions about whether α_{1M} is less than 0.64 would be considerably weaker under the other analyses. Differences are in part due to the treatment of typing errors. With an error rate of 0.01, the MCMC algorithm visits two modes: one where a captured parent is used and a typing error is invoked for the offspring; and another where an unobserved parent is used but no typing error is invoked. With an error rate of 0.05 (results not shown), the posterior mass shifted to the typing-error explanation and the parameters (α_{1M} in particular) were closer to the 'BY EYE' estimates. The other algorithms (as we used them) base their inferences on the single best family configuration (rather than considering multiple possibilities for parent assignments).

COLONY estimated a larger number of mothers per nest (i.e., λ was larger than by the other methods) and consequently underestimated the percentage of parents typed (i.e., g_0 was larger than by other methods). While none of the methods is a 'gold standard', examination of parent assignments shows that COLONY frequently assigned multiple parents when one parent could easily explain the data. The single parent was typically 'split' into multiple parents that are identical at most loci, but homozygous for different alleles at one or two loci, suggesting that the COLONY inferences were indeed overestimates for λ and g_0.

PARENTAGE also overestimated the proportion of uncaptured parents. This problem might be alleviated by a more sophisticated way of matching high posterior probability family configurations (as opposed to just the maximum a posteriori configuration) with the observed individuals, but this was not undertaken here. The high values of g_0 in both COLONY and PARENTAGE in turn influenced inferences for γ. Misidentifying individuals as 'uncaptured' seemed to pull the estimate of γ_1 upward, presumably making the value $\gamma_0 = \gamma_1 + \gamma_2$ closer to the observed fractions mothered by individuals inferred to be 'uncaptured'.

A major difference between the 'BYE EYE' and other approaches was the 'computation time' needed. The three computer based methods were comparable: COLONY required approximately 23 hours of computer time on a 1.6 Ghz Mac G5 Power PC; PARENTAGE required 31 hours; and our MCMC Bayesian required 37 hours on the same machine. The 'BY EYE' approach took about one month of investigator effort (although implementing a new computational method would have taken much longer).

DISCUSSION

Using our model, we were able to document intergroup (age, size, etc.) differences in reproductive success for a nest-guarding fish species. Age group affected maternal fertility in at least two ways: via the rate of nest participation, and via the proportion of eggs produced in nests with mothers from multiple age groups. The effect of age group on number of eggs is not modeled. This facilitates working with data where eggs have been sampled, and a total count may be unknown. It also eliminates the need to model the variability of nest size among nests with the same parental age make-up.

Bayesian inference applied to the model parameters showed that age is an important determinant of reproductive success in the mottled sculpin. Females appear to visit multiple males before spawning (Downhower and Brown 1979) and aquarium studies suggest that larger males are preferred by females (Brown and Downhower 1982). Furthermore, previous studies have shown evidence for positive size assortative mating in this species (Downhower *et al*. 1983; 1987). Because there is a general correlation in fishes between age and body size (Matthews 1998), we suspected that older males (and possibly older females) might be more successful in reproduction. Our results confirm that older individuals are more likely to contribute to nests (although for females, the differences between estimation methods indicates that interpretive caution is necessary). In addition, when females from different age groups spawned in the same nest, the older females contributed a higher proportion of the offspring. Grossman *et al*. (2002) previously showed that older female mottled sculpin have higher fecundities, based on dissection of gravid specimens. Our results demonstrate that this advantage in egg production carries over to the proportion of subsequently fertilized eggs that older-cohort females contribute to nests.

Our simulations show that increasing the number of nests improved the precision of all parameters estimates, and that such increases can largely compensate for low percentages of parents genotyped. In our case, increasing the number of offspring typed per nest (from 48 to 72) improved the estimates only of the parameters β and γ; however, more parameters would likely have been affected if the initial number of offspring had been inadequate to identify all parents contributing to the nest. If estimating the proportion of a nest contributed by mothers of each age group is of particular concern, more precise estimates could be efficiently obtained by augmenting the number of offspring genotyped for nests that already are identified as having two or more mothers from different age groups. Nests that satisfy this condition but that also have some mothers without age group data should, if possible, be avoided. Our computer simulations also demonstrate the effects of various sampling schemes on the precision of parameter estimates, but it is important to remember that the strategies considered for augmenting data will have different cost-utility trade offs for different organisms and different research questions.

For mottled sculpins, increasing the number of typed offspring per nest is relatively easy because most nests have large numbers of progeny. With the benefit of hindsight, if we had analyzed fewer progeny from each sculpin nest and increased the number of nests and adults assayed, we could have increased our power to detect differences in reproductive success between age groups with the same total genotyping effort. However, increasing the number of nests would have required sampling a larger stretch of stream, and increasing the percentage of parents typed would have been extremely difficult (because we already attempted to sample the population exhaustively). Increasing the number of sampled nests and parents might have also allowed us to increase the number of parameters that we estimated, and reduced the binning of age classes. This could certainly increase our understanding of the life-history of this species

and may have even increased our power to detect differences among the different age class if our choice of binning does not accurately reflect the biology of this species. Another option would be to follow Burzcky *et al.* (2006) and use simple parametric models for how nest participation and relative offspring production within a nest might vary with parental age.

Decreasing the number of loci that were analyzed could help to offset the costs associated with analyzing more nests. With five polymorphic loci, Fiumera *et al.* (2002) were able to reconstruct most of the parental genotypes 'BY EYE', and uniquely match these with adult genotypes in the population. With our method as applied to the sculpin data, reasonable inferences about many of the mating parameters could have been made with four or even three loci. For example, even for low numbers of loci, λ continued to correspond well to the values based on complete parentage information. This observation is consistent with the finding by DeWoody *et al.* (2000a, 2000b) that with merely two (highly polymorphic) loci, a sample of 48 offspring was often adequate to detect the number of distinct maternal parents in a half-sib family. Accurately matching typed adults in the population to nests does requires more loci and for parameters affected by these matches we see mild (α's, γ) to moderate (g_0) bias introduced when the number of loci is reduced to three.

Our method is an improvement over estimates derived using the COLONY maximum likelihood approach (Wang 2004), which overestimated the number of mothers per nest and the fraction of unobserved parents even with the full complement of five loci. This behavior is not affected by our post-hoc matching of inferred genotypes to observed adults, nor is it due to any failure to find the maximum likelihood configuration under the COLONY model. Under this model, additional parents are penalized only by a term representing their population genotype frequency. For large sibships and moderate numbers of loci, as considered here, this penalty is

frequently outweighed by an increased $P(O|M,F)$ for many offspring. Consequently, a multiple parent configuration for our large sibships frequently has a higher likelihood than a plausible single parent configuration, and COLONY will systematically fail to reconstruct a parsimonious assignment of parents. By using a multinomial model for the number of offspring belonging to each parent, our model discourages large differences between the proportions of offspring belonging to mothers in the same age group. The Poisson model for the number of mothers per nest also discourages large differences in the number of mothers across nests, even when different numbers of offspring are typed. This explicit modeling acts as an additional check on unnecessary splitting of sibships.

The program PARENTAGE was computationally efficient and performed well except in estimating the fraction of uncaptured parents. The difficulty in specifying priors for PARENTAGE's in-built mechanism for utilizing putative parent genotypes highlights a disadvantage of analyzing only one nest at a time; the algorithm cannot 'borrow strength' across nests to learn parameter values, so prior specification is more crucial. In this situation, use of a more flexible model for the number of offspring per parent, such as the dirichlet prior available in PARENTAGE, may not be an advantage. In addition, as we have implemented them, the posterior standard deviations for PARENTAGE (and COLONY) do not reflect uncertainty in parentage assignments, and will be underestimates when fewer genetic data are available. A more sophisticated method such as multiple imputation (Rubin 1987) may be able to use the PARENTAGE posterior samples to construct standard deviations that reflect this uncertainty.

Our MCMC Bayesian approach that explicitly accounts for nest structure in parentage analysis will likely find application to a variety of questions in evolutionary and conservation biology. One can imagine applying this approach to such cases as estimating the reproductive

success of wild versus hatchery released individuals (Dannewitz *et al.* 2004) or resident versus immigrants (Johannesen and Andreassen 1998). It is important to remember that parentage analyses require extensive genotyping; careful consideration should be taken to ensure that adequate sample sizes of nests, offspring per nest, and parents can be obtained to allow robust parameter estimates.

ACKNOWLEDGEMENTS

Thanks go to J. Wang for assistance in installing and running COLONY, to I. J. Wilson for assistance in running PARENTAGE, and to the referees and associate editor for insightful comments. This work was supported by by a Fast Start Grant from the Marsden fund of New Zealand (BJ) a National Institutes of Health Ruth L. Kirchstein Postdoctoral Fellowship (ACF), and grants to GDG from the U.S.D.A. Forest Service McIntire-Stennis program (GEO-0086-MS), National Science Foundation (DEB-2018001), and Warnell School of Forestry and Natural Resources.

LITERATURE CITED

AVISE, J.C., A.G. JONES, D. WALKER, J.A. DEWOODY, B. DAKIN, A. FIUMERA, *et al.*, 2002 Genetic mating systems and reproductive natural histories of fishes: lessons for ecology and evolution. Annu. Rev. Genet. **36:** 19-45.

ADAMS, W. T., A. R. GRIFFIN and G. F. MORAN, 1992 Using paternity analysis to measure effective pollen dispersal in plant populations. Am. Nat. **140:** 762-780.

BIRKHEAD, T. R., and A. P. MØLLER, 1995 Extra-pair copulation and extra-pair paternity in birds. Anim. Behav. **49:** 843-848.

BROWN, L., and J. F. DOWNHOWER, 1982 Polygamy in the Mottled Sculpins (*Cottus, bairdi*) of Southwestern Montana (Pisces, Cottidae). Can. J. Zool. **60:** 1973-1980.

BURCZYK, J., W. T. ADAMS, D. S. BIRKES and I. J. CHYBICKI, 2006 Using genetic markers to directly estimate gene flow and reproductive success parameters in plants on the basis of naturally regenerated seedlings. Genetics **173:** 363-372.

BUTLER, K., C. FIELD, C. M. HERBINGER and B. R. SMITH, 2004 Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. Mol. Ecol. **13:** 1589-1600.

CURRY, R. A., 2005 Assessing the reproductive contributions of sympatric anadromous and freshwater-resident brook trout. J. Fish Biol. **66:** 741-757.

DANNEWITZ, J., E. PETERSSON, J. DAHL, T. PRESTEGAARD, A. C. LOF *et al.*, 2004 Reproductive success of hatchery-produced and wild-born brown trout in an experimental stream. J. Appl. Ecol. **41:** 355-364.

DEWOODY, J. A., Y. D. DEWOODY, A. C. FIUMERA and J. C. AVISE, 2000a On the number of reproductives contributing to a half-sib progeny array. Genet. Res. **75:** 95-105.

DEWOODY, J. A., D. E. FLETCHER, S. D. WILKINS and J. C. AVISE, 2000 Parentage and nest guarding in the Tessellated Darter (*Etheostoma olmstedi*) assayed by microsatellite markers (Perciformes : Percidae). Copeia**:** 740-747.

DOWNHOWER, J. F., and L. BROWN, 1979 Seasonal-Changes in the Social-Structure of a Mottled Sculpin (*Cottus-bairdi*) Population. Anim. Behav. **27:** 451-458.

DOWNHOWER, J. F., L. BROWN, R. PEDERSON and G. STAPLES, 1983 Sexual selection and sexual dimorphism in mottled sculpins. Evolution **37:** 96-103.

DOWNHOWER, J. F., L. S. BLUMER and L. BROWN, 1987 Seasonal variation in sexual selection in the mottled sculpin. Evolution **41:** 1386-1394.

EMERY, A. M., I. J. WILSON, S. CRAIG, P.R. BOYLE and L.R. NOBLE, 2001 Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. Mol. Ecol. **10**: 1265-1278.

FIUMERA, A. C., B. A. PORTER, G. D. GROSSMAN and J. C. AVISE, 2002 Intensive genetic assessment of the mating system and reproductive success in a semi-closed population of the mottled sculpin, *Cottus bairdi*. Mol. Ecol. **11:** 2367-2377.

GERBER, S., P. CHABRIER and A. KREMER, 2003 FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. Mol. Ecol. Notes **3:** 479-481.

GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711-732.

GROSSMAN, G. D., K. MCDANIEL and R. E. RATAJCZAK, 2002 Demographic characteristics of female mottled sculpin, *Cottus bairdi*, in the Coweeta Creek drainage, North Carolina. Environ. Biol. Fish. **63:** 299-308.

HASTINGS, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:**97-109.

JOHANNESEN, E., and C. P. ANDREASSEN, 1998 Survival and reproduction of resident and immigrant female root voles (Microtus oeconomus). Can. J. Zool. **76:** 763-766.

JONES, B., 2003 Maximum likelihood inference for seed and pollen dispersal distributions. J. Agric. Biol. Envir S. **8:** 170-183.

KING, R. B., W. B. MILSTEAD, H. L. GIBBS, M. R. PROSSER, G. M. BURGHARDT *et al.*, 2001 Application of microsatellite DNA markers to discriminate between maternal and genetic effects on scalation and behavior in multiply-sired garter snake litters. Can. J. Zool. **79:** 121-128.

LEVITAN, D. R., 2005 The distribution of male and female reproductive success in a broadcast spawning marine invertebrate. Integr. Comp. Biol. **45:** 848-855.

MARSHALL, T. C., J. SLATE, L. E. B. KRUUK and J. M. PEMBERTON, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7:** 639-655.

MATTHEWS, W.J., 1998 Patterns in Freshwater Fish Ecology. Chapman and Hall, New York, 731. pp.

MYERS, E. M., and K. R. ZAMUDIO, 2004 Multiple paternity in an aggregate breeding amphibian: the effect of reproductive skew on estimates of male reproductive success. Mol. Ecol. **13:** 1951-1963.

NASON, J. D., E. A. HERRE and J. L. HAMRICK, 1998 The breeding structure of a tropical keystone plant resource. Nature **391:** 685-687.

NEFF, B. D., J. REPKA and M. R. GROSS, 2000 Parentage analysis with incomplete sampling of candidate parents and offspring. Mol. Ecol. **9:** 515-528.

NIELSEN, R., D. K. MATTILA, P. J. CLAPHAM and P. J. PALSBOLL, 2001 Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. Genetics **157:** 1673-1682.

RØED, K. H., O. HOLAND, H. GJOSTEIN and H. HANSEN, 2005 Variation in male reproductive success in a wild population of reindeer. J. Wildlife. Manage. **69:** 1163-1170.

ROEDER, K., B. DEVLIN and B. G. LINDSAY, 1989 Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. Biometrics **45:** 363-379.

RUBIN, D. B., 1987 Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

SAVAGE, T., 1963 Reproductive behavior in the mottled sculpin, *Cottus bairdi* Girard. Copeia **1963:** 317-325.

SHURTLIFF, Q. R., D. E. PEARSE and D. S. ROGERS, 2005 Parentage analysis of the canyon mouse (*Peromyscus crinitus*): Evidence for multiple paternity. J. Mammal. **86:** 531-540.

SIEBERTS, S. K., E. M. WIJSMAN and E. A. THOMPSON, 2002 Relationship inference from trios of individuals, in the presence of typing error. Am. J. Hum. Genet. **70:** 170-180.

SLATE, J., P. M. VISSCHER, S. MACGREGOR, D. STEVENS, M. L. TATE *et al.*, 2002 A genome scan for quantitative trait loci in a wild population of red deer (*Cervus elaphus*). Genetics **162:** 1863-1873.

SMOUSE, P. E., and T. R. MEAGHER, 1994 Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L) Gray (Liliaceae). Genetics **136:** 313-322.

Walker, D., A.J. Power, M. Sweeney-Reeves, and J.C. Avise, 2007 Multiple paternity and female sperm usage along egg-case strings of the knobbed whelk, *Busycon carica* (Mollusca; Melongenidae). Marine Biol. **151:** 53-61.

WANG, J. L., 2004 Sibship reconstruction from genetic data with typing errors. Genetics **166:** 1963-1979.

APPENDIX: MARKOV CHAIN MONTE CARLO ALGORITHM

The algorithm is initialized by assigning each nest a single 'unobserved' father whose genotype at each locus consists of the two most frequent alleles observed among the offspring at that locus. Conditional on this father, for each offspring, the observed mother maximizing the probability of that offspring's genotype is then chosen and added to the nest. Because typing error is possible, there is always a mother resulting in a non-zero probability for the offspring. The parameters are initialized at a configuration that encourages a relatively parsimonious assignment of parents (p=0.95, λ=4.0); other parameters are set at their prior mean. The nest configuration is then updated for 10,000 'burn-in' steps before the parameters are updated. An update of the nest configuration consists of updating each nest by proposing one of the following moves (where necessary the sex of the parent to be updated is also selected at random, with each sex picked with probability 0.5):

1. Add an unobserved parent. The genotype of the new unobserved parent is constructed by selecting two offspring, with probability inversely proportional to their genotype's probability under the best pair from among the parents currently assigned to the nest. However, a lower bound of 0.0001 is placed on an offspring's probability; otherwise, offspring with typing errors would be picked almost exclusively. Then, for each locus an allele is randomly selected from each offspring to construct the new parent's genotype. A limit is set of 10 unobserved parents of each gender per nest; proposals to add above this limit are automatically rejected. Addition of an unobserved parent changes the dimension of the unobserved quantities we are sampling over by adding an unknown genotype. The algorithm is in fact a reversible jump algorithm (Green, 1995); however, because the additional parameters are discrete, the relevant Jacobian is 1.0 and there is no difference from the 'ordinary' Hastings ratio.

2. Add an observed parent. The proposed parent is picked at random from among observed parents of the selected sex not yet assigned to the nest.

3. Swap an observed parent for another observed parent. The parent to be swapped out is selected at random from those currently assigned to the next with the selected sex; the parent to be swapped in is selected at random from observed parents of the selected sex not yet assigned to the nest. This move and move (4) are automatically rejected if there are no current observed parents of the selected sex.

4. Swap a current observed parent for a new unobserved one. The genotype of the new unobserved parent is proposed as in (1), the parent to be replaced is selected at random from the current observed parents of the selected sex. This move is automatically rejected if it results in more than 10 unobserved parents of one sex.

5. Swap a current unobserved parent for an observed one. The new observed parent is selected at random from observed parents of the selected sex not already assigned to the nest.

6. Delete a parent. Select at random from among parents of the selected sex. The move is automatically rejected if it would leave the nest with no parents of one sex.

7. Swap the primary father with cuckolding father. The cuckolder is randomly selected from among all cuckolders; the move is automatically rejected if there are no cuckolders.

8. Update the genotype of an unobserved parent. This is similar to the procedure outlined in (1), except that only one offspring is picked and only one allele at each locus is updated.

For each move, the Hastings ratio, a ratio of the posterior densities and proposal probabilities, is calculated; the move is accepted with probability min(1, Hastings ratio). After the first 10,000 iterations, the parameters are updated after every 10 nest configuration updates, and recorded every 500 such updates. Most parameters are updated using a random walk proposal—a small

change in the parameter is proposed, and accepted or rejected using the Hastings ratio as outlined above. It is possible to propose from the full conditional posterior distributions of α_F and g (both dirichlet distributions); this guarantees a Hastings ratio of 1.0. In total, we conducted 2.5 million nest configuration updates, resulting in 49,800 samples of the parameters.

Table 1. Description of parameters estimated.

Parameter	Value used in simulations	Definition of parameter
λ	2.87	The total number of mothers participating in a nest is a truncated Poisson with this parameter. The mean number of mothers per nest is $\lambda/\{1-exp(-\lambda)\}$.
p	0.96	Defines the number of fathers in a nest given a geometric distribution. The mean number of fathers is $1/p$.
α_{iM}	0.58	The probability that a mother in a nest is from group *1*.
α_{iF}	0.57	The probability that a father in a nest is from group *1*.
γ_i	0.36	Governs the fraction of offspring produced by mothers in age class 1 conditional on nest participation from multiple groups.
β	0.31	Probability that an offspring in a nest is sired by a cuckolding father.
g_0	0.5	Proportion of the parent population that has not been sampled.

Table 2. Comparison of results for the Markov Chain Monte Carlo, 'BYE EYE', COLONY, and PARENTAGE methods for the mottled sculpin data.

Parameter	λ	α_{1M}	γ_1	p	α_{1F}	β	g_0
MCMC Mean, *sd*	2.84, *0.39*	0.52, *0.07*	0.34, *0.02*	0.84, *0.08*	0.57, *0.13*	0.35, *0.06*	0.47, *0.05*
'BY EYE' Mean, *sd*	2.86, *0.38*	0.62, *0.07*	0.37, *0.02*	0.92, *0.05*	0.56, *0.12*	0.46, *0.05*	0.51, *0.05*
COLONY Mean, *sd*	5.23, *0.48*	0.67, *0.12*	0.47, *0.04*	0.86, *0.06*	0.57, *0.12*	0.26, *0.02*	0.71, *0.04*
PARENTAGE Mean, *sd*	2.96, *0.38*	0.65, *0.11*	0.48, *0.04*	0.92, *0.05*	0.70, *0.14*	0.32, *0.06*	0.73, *0.04*

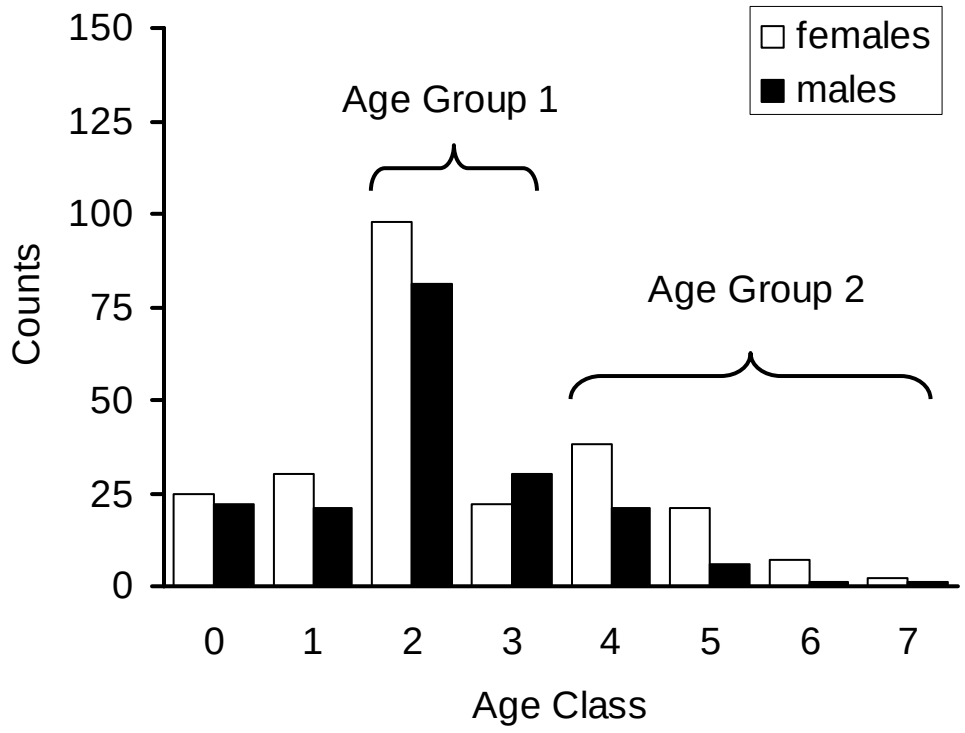
FIGURE LEGENDS

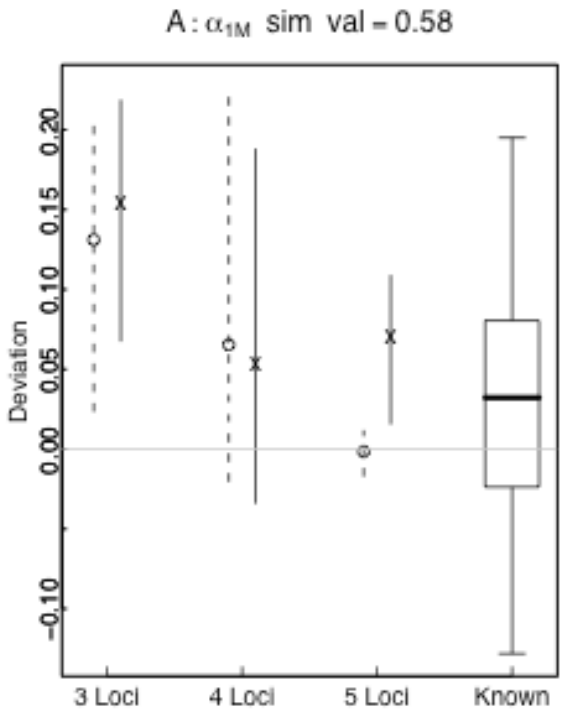
Figure 1. Histogram of age classes. The age class distributions for females (white bars) and males (black bars) are shown separately for 243 female and 186 male *Cottus bairdi* that were successfully aged and sexed. The binnings into age group 1 and age group 2 are shown.

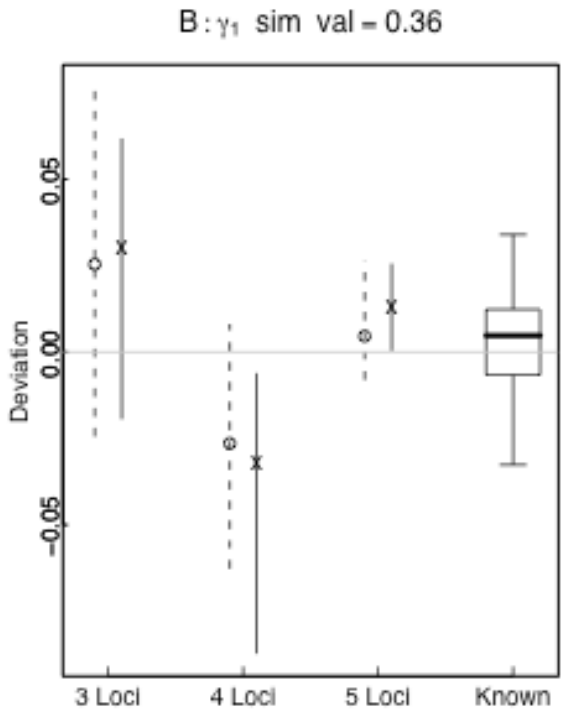
Figure 2. Deviations between the MCMC and parentage known estimates, and between the MCMC estimates and simulation values, for the parameters measuring differences between age groups ($\alpha_{1M,}$ $\gamma_{1,}$ α_{1F}). Dashed lines represent the range of deviations between the MCMC estimates and parentage known estimates, with the mean deviation given by a circle, over five replicate simulations for each of three, four, and five loci. Solid lines give the corresponding range of deviations between the MCMC estimate and the simulation values, with the mean deviation given by an x. The horizontal grey line indicates zero deviation. The boxplot shows deviations between the parentage known estimates and simulation values for all 15 simulations.

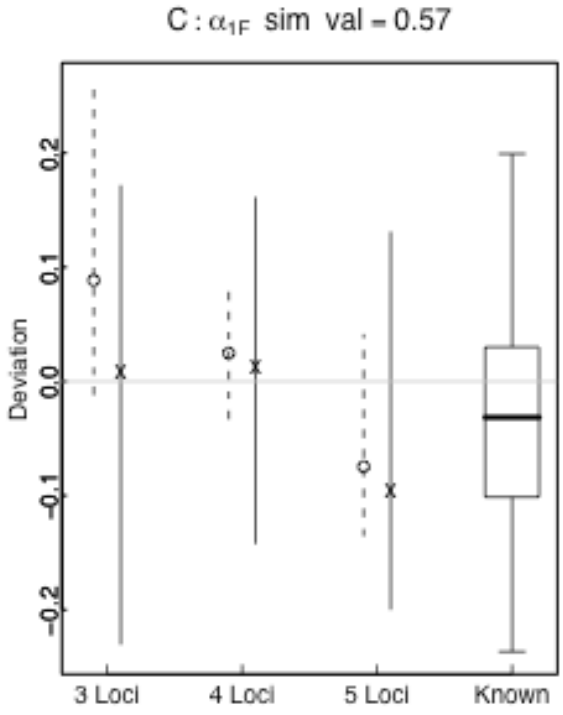
Figure 3. Deviations between the MCMC and parentage known estimates, and between the MCMC estimates and simulation values, for the parameters λ, p, β, g_0. Dashed lines represent the range of deviations between the MCMC estimates and parentage known estimates, with the mean deviation given by a circle, over five replicate simulations for each of three, four, and five loci. Solid lines give the corresponding range of deviations between the MCMC estimate and the simulation values, with the mean deviation given by an x. The horizontal grey line indicates zero deviation. The boxplot shows the deviations between the parentage known estimates and simulation values for all 15 simulations.

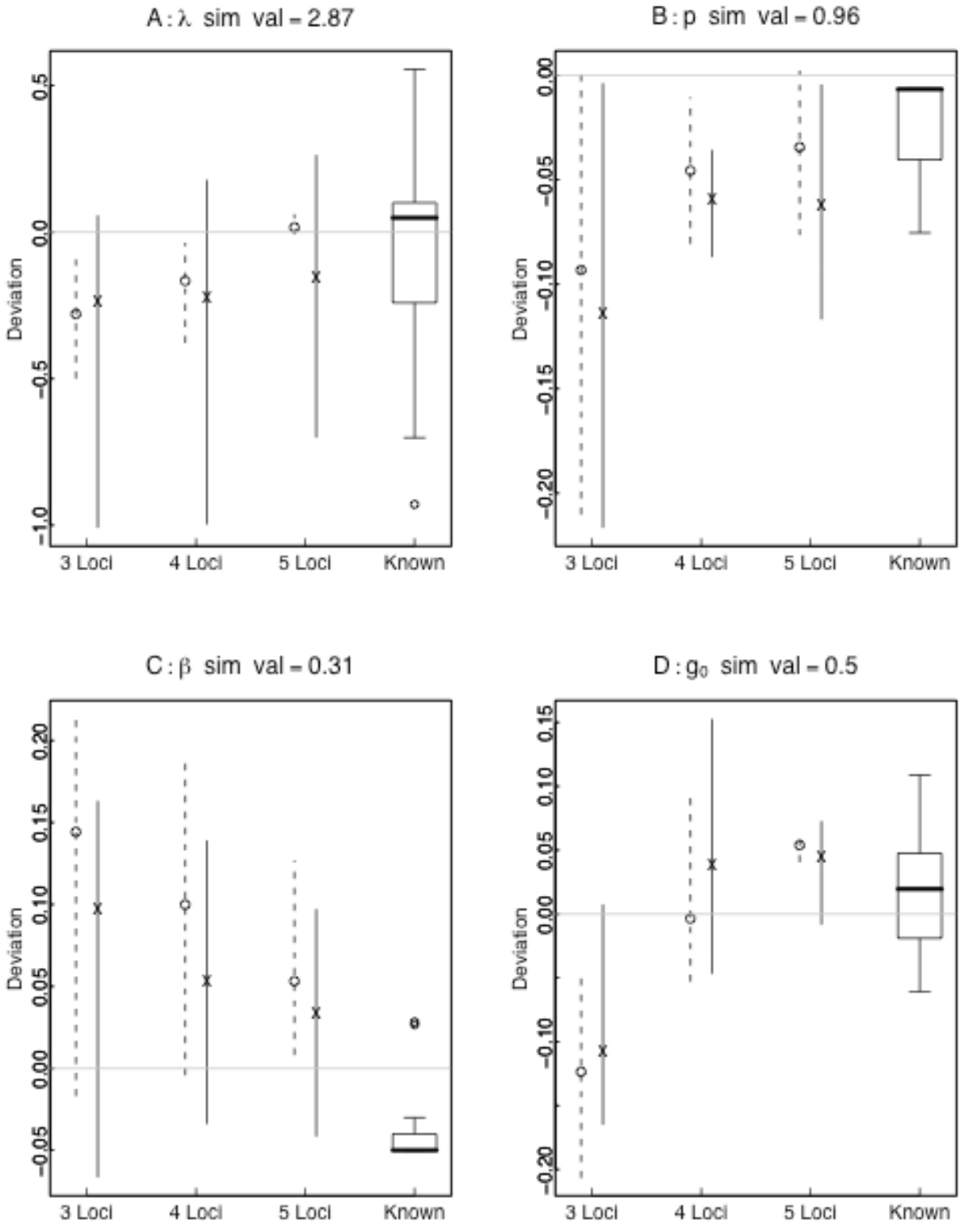
Figure 4. Effects of varying data collection scenarios on the posterior standard deviations of the parameter estimates. The effects of increasing the proportion of genotyped parents (to 75%), the number of analyzed nests (to 33), the number of offspring (to 72), or analyzing 44 nests but only sampling 25% of the genotyped parents, are shown for: (A) age group parameters, and (B) other parameters. The solid line corresponds to no change; the dashed line indicates a decrease in the standard deviation by $1/\sqrt{1.5}$; and the dotted line indicates a decrease in the standard deviation by $1/\sqrt{2}$.

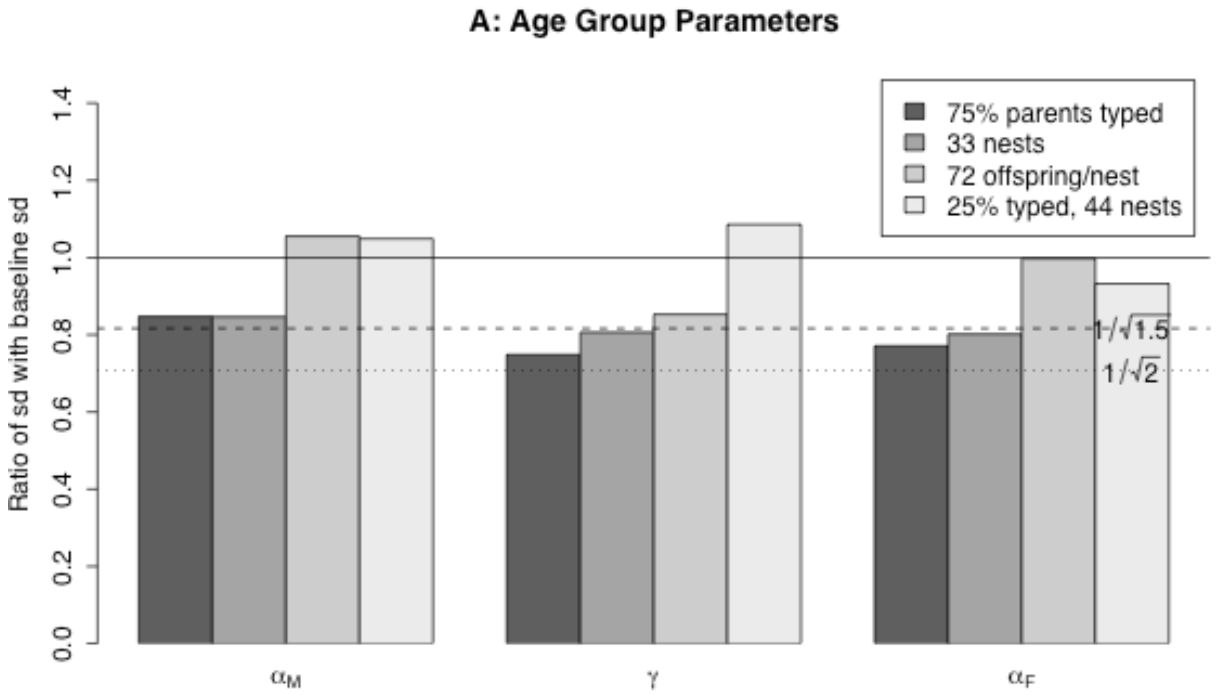

150
125
100
75
50
25
0
Counts
females
males
Age Group 1
Age Group 2
0
1
2
3
4
5
6
7
Age Class

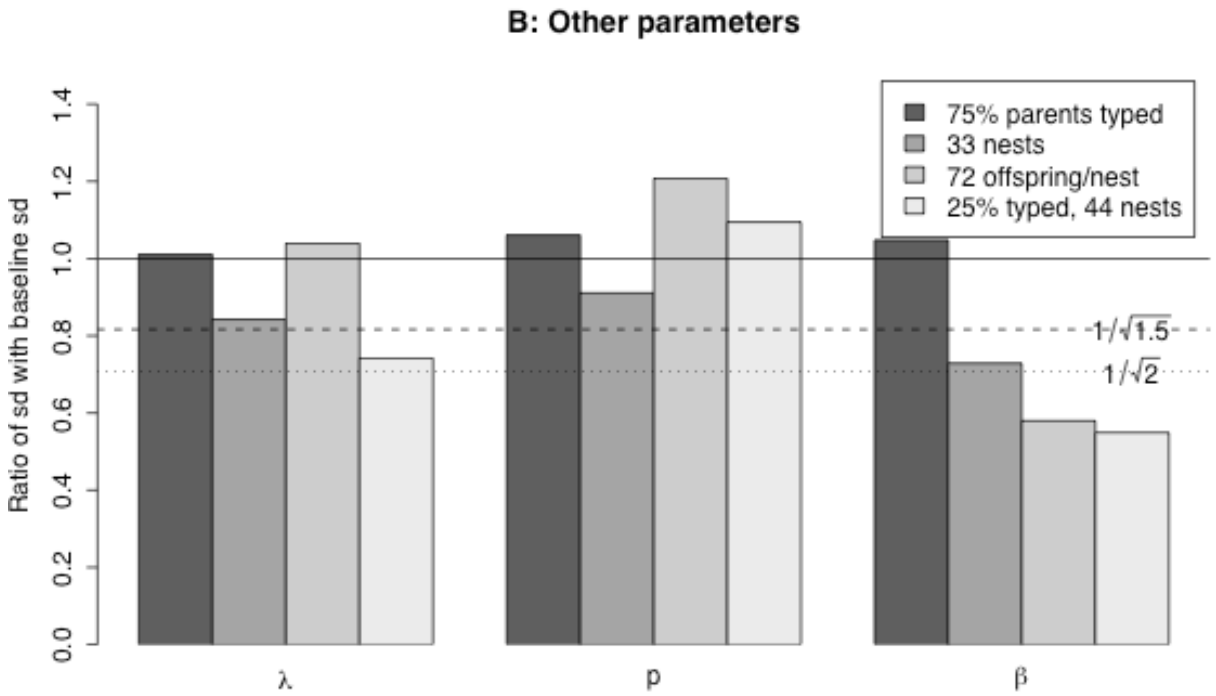

A : α_{1M} sim val = 0.58
Deviation
0.20
0.15
0.10
0.05
0.00
−0.10
3 Loci
4 Loci
5 Loci
Known


B : γ_1 sim val = 0.36
Deviation
0.05
0.00
−0.05
3 Loci
4 Loci
5 Loci
Known


C : α_{1F} sim val = 0.57
Deviation
0.2
0.1
0.0
−0.1
−0.2
3 Loci
4 Loci
5 Loci
Known


A : λ sim val = 2.87
B : p sim val = 0.96
C : β sim val = 0.31
D : g0 sim val = 0.5
Deviation
3 Loci
4 Loci
5 Loci
Known


A: Age Group Parameters
Ratio of sd with baseline sd
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
75% parents typed
33 nests
72 offspring/nest
25% typed, 44 nests
1/√1.5
1/√2
αM
γ
αF


B: Other parameters
Ratio of sd with baseline sd
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
75% parents typed
33 nests
72 offspring/nest
25% typed, 44 nests
1/√1.5
1/√2
λ
p
β