Genet. Res., Camb. (2000), **75**, *pp.* 95–105. *With* 9 *figures.* Printed in the United Kingdom © 2000 Cambridge University Press

On the number of reproductives contributing to a half-sib progeny array

J. ANDREW DEWOODY*, YSSA D. DEWOODY†, ANTHONY C. FIUMERA AND JOHN C. AVISE

Departments of Genetics and of Mathematics†, University of Georgia, Athens, GA 30602, USA

(*Received 29 September 1998 and in revised form 5 January 1999*)

Summary

We address various statistical aspects of biological parentage in multi-offspring broods that arise via multiple paternity or multiple maternity and, hence, consist of mixtures of full- and half-sibs. Conditioned on population genetic parameters, computer simulations described herein permit estimation of: (1) the mean number of offspring needed to detect all parental gametes in a brood and (2) the relationship between the number of distinct parental gametes found in a brood and the number of parents. Results are relevant to the design of empirical studies employing molecular markers to assess genetic parentage in polygynous or polyandrous species with large broods, such as are found in many fishes, amphibians, insects, plants and other groups. The utility of this approach is illustrated using two empirical data sets.

1. Introduction

Many animal and plant species with polygamous mating systems may produce individual broods that consist of a mixture of full-sib and half-sib offspring. In such cases, the clutch may be the product of multiple paternity (e.g. Cobbs, 1977; Griffiths *et al.*, 1982; Baker *et al.*, 1999) or multiple maternity (e.g. Jones & Avise, 1997*a*; DeWoody *et al.*, 1998). Biologists are interested in the number of parents and their relative contributions to half-sib progeny arrays for several reasons, including assessments of multiple insemination (Levine *et al.*, 1980), brood parasitism and other reproductive behaviours (Jones & Avise, 1997*a*, *b*; Coltman *et al.*, 1998; DeWoody *et al.*, 1998; Fitzsimmons, 1998; Imhof *et al.*, 1998; Jones *et al.*, 1998; Kellogg *et al.*, 1998; Baker *et al.*, 1999). However, challenging statistical and sampling issues arise in using molecular genetic markers to estimate parental contributions to a half-sib brood.

A half-sib clutch may be evidenced if, in the progeny array, more than four alleles are present at any one locus (Levine *et al.*, 1980). In a typical molecular analysis of a potential half-sib clutch, gametic genotypes or haplotypes ('gametotypes') tracing to the unshared parent(s) can be deduced by subtracting each progeny's diploid genotype from that of the known parent. In a single-locus assessment (and barring *de novo* mutation), the *minimum* number of unshared parents who contributed to a half-sib brood is simply the smallest integer value greater than or equal to one-half the number of different gametotypes inherited by progeny from those unshared parents (Kellogg *et al.*, 1998). The difference between this minimal estimate and the true number of unshared parents is some function of how often adults share alleles. Normally, the two values are expected to be identical only in hypothetical cases where each allele in the parental population is unique. Some microsatellite loci may approach this ideal, but even highly polymorphic markers fall short of overcoming the limitations of face-value empirical estimates of parental contributions to a clutch. Thus, a remaining question is how many parents actually contributed to a progeny array.

Alleles shared among parents can also complicate decisions about how many progeny must be sampled from a half-sib cohort to detect all parental gametes present. Family sizes in many insects, amphibians and plants are far larger than can be analysed feasibly in the laboratory (Nason *et al.*, 1996; Fletcher *et al.*, submitted). In many fish, for example, several thousand embryos may be present in a single nest (Taborsky, 1994). How many of these should be

* Corresponding author. Tel: +1 (706) 542 1448. Fax: +1 (706) 542 3910. e-mail: dewoody@arches.uga.edu.

sampled to provide reasonable assurance of detecting all parents who contributed to the brood?

Likelihood-based methods for estimating re-mating frequency and sperm displacement have recently been developed for *Drosophila* (Harshman & Clark, 1998), but these authors assume a geometric decline in the relative contributions of each successive unshared parent. This assumption may not be valid in many organisms with external fertilization. Other programs that 'assign' putative parents to a brood (e.g. Smouse & Meagher, 1994; Marshall *et al.*, 1998) require genotypes of each potential parent. We have developed simulation programs that, in the absence of exhaustive genotypic data, allow estimation of (1) the number of parents that contributed to a half-sib progeny array and (2) the number of offspring that must be sampled from an array to detect at least one gamete from each parent.

Table 1. *Parameter variables that can be specified for the programs BROOD, HAPLOTYPES and GAMETES*

BROOD	HAPLOTYPES/ GAMETES
Number of loci	Number of loci
Number of alleles, allele frequencies	Number of alleles, allele frequencies
Adult population size	Adult population size
Number of unshared parents	Maximum number of unshared parents
Size of progeny array	Size of progeny array
Number of times parents re-sampled	Number of times parents re-sampled
Number of times progeny re-sampled	Number of times progeny re-sampled
Relative contributions of unshared parents	Number of progeny sampled

2. Materials and methods

The first part of this section will outline the general methods underlying our models.

Suppose one has reasonable empirical estimates of gene frequencies at various neutral loci within a population. In principle, one can then use these frequencies to randomly generate 'individuals' whose genotypes are products of Hardy–Weinberg equilibrium. Pairs of these individuals can then be randomly drawn from the population and 'mated' such that the resulting full-sib progeny arrays are products of conventional Mendelian inheritance. If half-sib progeny arrays are desired, one must add the parameter of reproductive skew (i.e. the proportion of the shared parent's offspring which stem from each unshared parent).

We used this logic to devise computer programs which address the two questions outlined in Section 1. Each program repeatedly generates half-sib broods in which diploid genotypes of all sampled offspring and their one shared parent are known, but genotypes of unshared parents are not required (Table 1; Fig. 1).

(i) *Sampling regimes*

The first program calculates two statistics and their associated variance. These two statistics are monitored one locus at a time, and the most informative locus from a suite of loci is used to determine appropriate sample sizes from the progeny arrays. The first statistic, n, is the number of offspring needed per clutch to detect all marker-unique gametes from the unshared parents of the clutch (Fig. 1). That is, if four distinct alleles at a locus are present among three unshared parents, $\bar{n}$ is the mean sample size from the clutch needed to detect all four of these parental alleles. Note in this case that not all alleles are unique in state (i.e. only four alleles are represented among the six parental chromosomes). The second statistic, n^*, attempts to account for this problem of non-unique alleles. The parameter n^* is the number of offspring per clutch needed to observe all *true* gametotypes (not merely those detected by available markers) from unshared parents of the clutch.

Thus, the *minimum* value of n is equal to the number of distinct alleles in a brood contributed by unshared parents at the most polymorphic locus, whereas the minimum value of n^* is equal to twice the number of unshared parents. For example, if two mothers of genotype AB and BC contribute to a half-sib progeny array, the minimum value of n is 3, whereas the minimum value of n^* is 4. It should be clear that n will always be less than or equal to n^*.

Distributions of n and n^* values were generated via computer by sampling from a half-sib progeny array hundreds or thousands of times. By sampling new progeny arrays (created by re-sampling parents from the initial adult population), each resulting brood is independent and statistics can be averaged across arrays. Thus, sampling issues are examined both within and among broods. These simulated sample sizes ($\bar{n}$ and $\bar{n}^*$) can then be used as guidelines for empirical sample sizes of progeny.

(ii) *Parental assemblage size*

Assuming that half-sib progeny arrays can be generated in the manner described above, gametotypes can in principle be used to determine the number of parents. Repeated sampling of progeny arrays generated by a fixed number of parents should produce a distribution of the number of distinct gametotypes represented in each array. For example, if six females

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

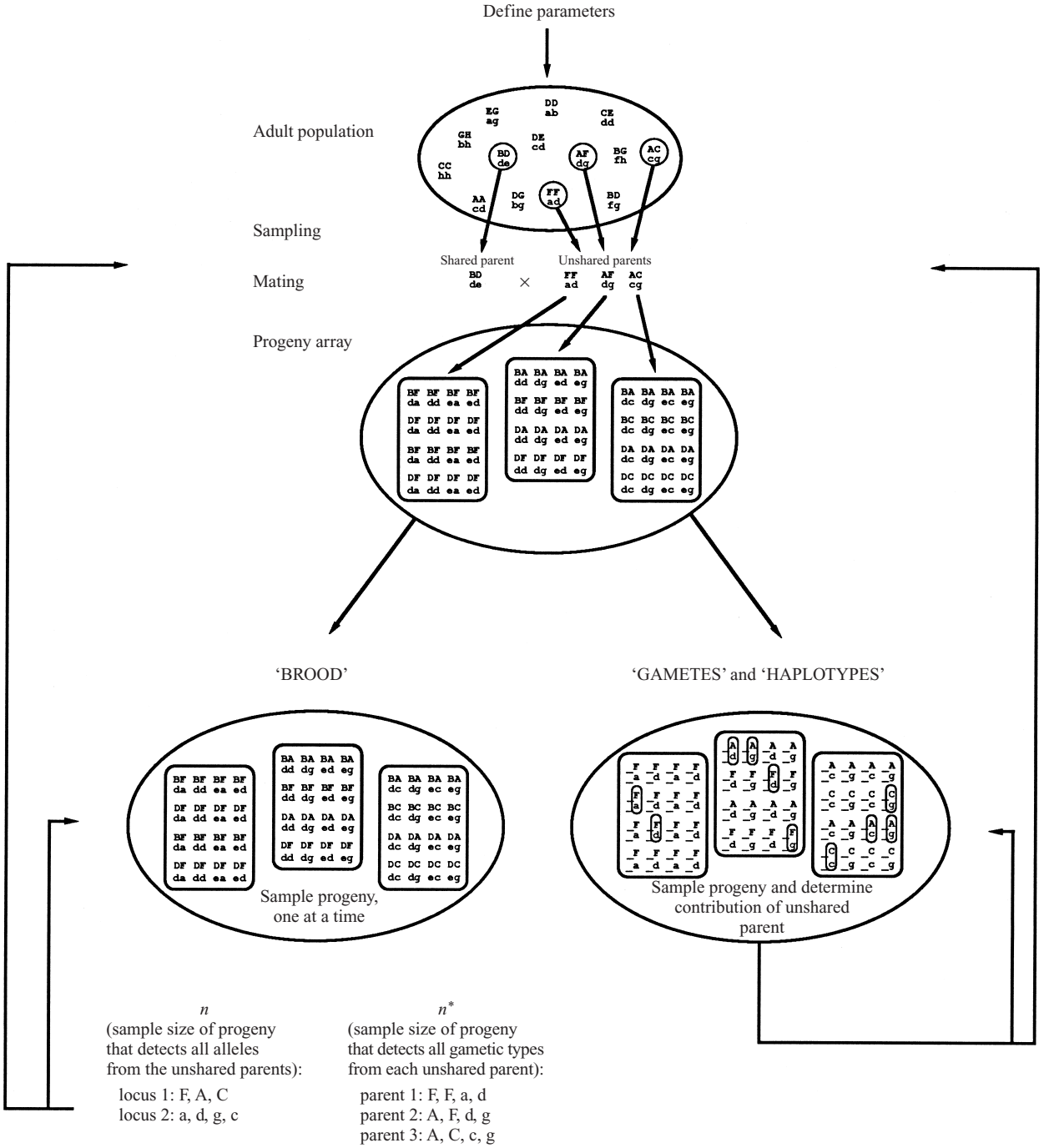

Fig. 1. Flowchart illustrating the logic underlying the simulation procedures. Initially, the user defines certain population genetic parameters (Table 1) that characterize the adult population. The shared and unshared parents are randomly chosen from the adult population, and progeny arrays (broods or clutches) are then created. Diploid genotypes are shown at two unlinked loci (upper- and lower-case letters, respectively). BROOD reports the mean, variance and confidence intervals around the number of progeny that must be sampled to detect the genetic contributions of each parent (see text). GAMETES and HAPLOTYPES utilize knowledge of the shared parent's genotype to deduce the genetic contributions (i.e. 'gametotypes') of the unshared parents. The number of gametotypes contributed by parental assemblages of various size are then used to estimate the number of parents that contributed to a progeny array (see text).

contribute equally to a singly-sired brood, simulations may determine that 95% of the time between eight and 10 gametotypes are detected at the most polymorphic locus. These distributions can then be constructed for various parental 'assemblage' sizes, where each assemblage consists of the one shared parent and x unshared parents, where x assumes all integer values between 1 and some explicit maximum.

Once gametic distributions are generated, the process can be inverted such that assemblage size now

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

Table 2. *Parameters used for the simulations described in the text*

Parameter	High polymorphism	Low polymorphism
Number of loci	2	2
Number of alleles	25 and 15	5 at each locus
Allele frequencies	(25 alleles) 15 @ 0·053, 10 @ 0·020 (15 alleles) 9 @ 0·089, 6 @ 0·033	3 @ 0·2667, 2 @ 0·1
Maximum assemblage size	15	15
Adult population size	500	500
Size of progeny array	500	500
Number of times parents re-sampled	1000	1000
Number of times progeny re-sampled	1000	1000
Number of progeny sampled	50 for GAMETES, 100 for HAPLOTYPES	50 for GAMETES, 100 for HAPLOTYPES

is dependent on the number of differentiable gametotypes found in the samples. This inversion process is accomplished by pooling all the above distributions and then sorting the individual experiments by the number of gametotypes detected. Based on the new distributions, a mean assemblage size of parents and a 95% confidence interval about it are associated with the number of different gametotypes identified in a brood sample. For example, if there are 12 different maternal gametes within a half-sib brood, we might determine with 95% confidence that there are between six and nine mothers.

Imagine a species where one sex (say the male) mates with up to 15 individuals of the opposite sex, each of whom is parent to an equal number of progeny in a brood. Fifteen different assemblages then are created: one male mated with one female, with two females, with three females, and so on up to one male mated with 15 females. The parents in each assemblage are chosen at random from the adult population and a brood is created. A sample from this brood is analysed and the number of distinct gametes recorded. New parents then are sampled from the adult population and the process repeated with the same assemblage size. Finally, the entire protocol is repeated for each different assemblage size.

If specific distributions are used to define the contributions of unshared parents to the brood (i.e. the reproductive skew), the entire discrete parameter space can in principle be explored through simulations. Such an extension would provide a maximum-likelihood estimate of the true parental assemblage size as determined by Monte Carlo simulations.

(iii) *Simulation parameters*

We first tested the relative importance of various biological parameters on the sample sizes of offspring (n and n^*) necessary to detect distinct parental gametes in a brood. First, the numbers of unshared parents and the relative contributions of each were varied while the number of loci, alleles and allele frequencies were held constant. Next, the number of unshared parents and their relative contributions were held constant while the numbers of alleles and their relative frequencies were altered. Then, numbers of alleles and of loci were varied to determine their effects on n and n^*.

The level of polymorphism was adjusted by varying the number of alleles per locus and the allele frequency distributions. Allele frequency distributions were

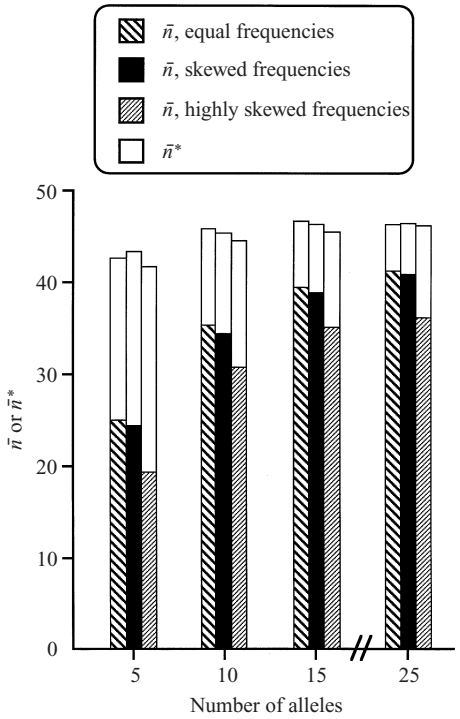

Fig. 2. Examples of the effects of the number of alleles and allele frequencies at a single locus on $\bar{n}$ and $\bar{n}^*$ when the number of unshared parents (4) and the relative contributions of each (7:1:1:1) were held constant. Forty per cent of the alleles contributed equally to 80% of the gametic pool and the other alleles contributed equally to the remaining 20%. For example, in the five-allele case, two alleles each have a frequency of 0·4 and the other three alleles have a frequency of 0·0667 each.

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

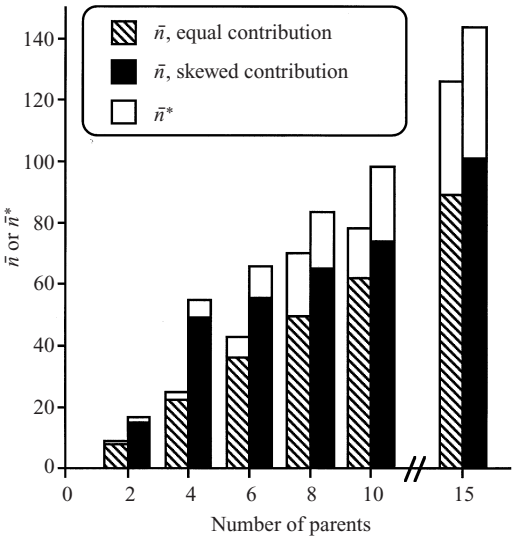

Fig. 3. Examples of the effects of the number of unshared parents and the relative contribution of each on $\bar{n}$ and $\bar{n}^*$ when the number of loci (2), number of alleles (25 and 15, respectively) and allele frequencies were held constant, such that 40% of the alleles contributed equally to 80% of the gametic pool (see text). 'Contribution' refers to the fraction of the brood attributable to the most successful unshared parent; other unshared parents contributed equally to the remainder of the brood.

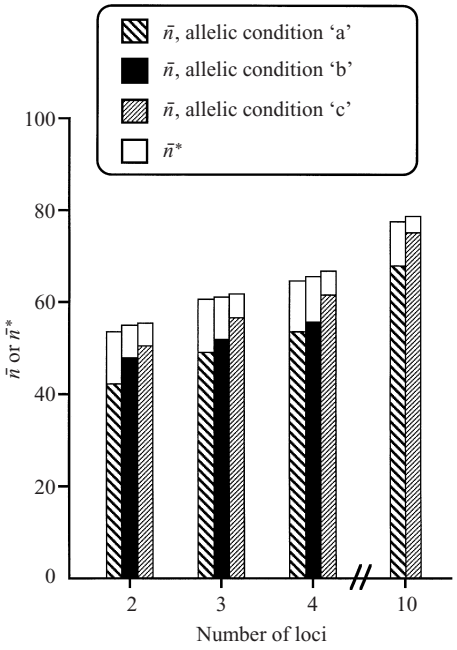

Fig. 4. Examples of the effects of allelic polymorphism and number of loci on $\bar{n}$ and $\bar{n}^*$ when the number of unshared parents (4) and their relative contributions to a brood were held constant. One parent contributes 70% of the fertilizations and the other three parents contribute 10% each. Allele frequencies are skewed such that 60% of the alleles contributed equally to 80% of the gametic pool and the other alleles contributed equally to the remaining 20%. Allelic condition 'a' refers to 10 alleles-per-locus ('apl') in all cases. Condition 'b' refers to 15 apl in the two-locus case, 15/15/10 apl in the three-locus case and 15/15/10/10 apl in the four-locus case. Condition 'c' refers to 25 apl in all cases.

designated such that a fraction of the alleles at a locus contributed equally to a fraction of the gametic pool while the remaining alleles contributed equally to the remainder of the gametic pool. For example, if 40% of the alleles contributed equally to 80% of the gametic pool (and the other alleles contributed equally to the remaining 20%), then for the five-allele case, two alleles each have a frequency of 0·4 and the other three alleles have a frequency of 0·0667 each.

A second series of simulations under specified conditions (Table 2) was used to estimate numbers of unshared parents that contributed to a brood. Distributions were then constructed showing the relationship between numbers of unshared parents and the observed numbers of gametes or haplotypes given these population genetic conditions. These distributions were then pooled and inverted as described earlier, thus creating new distributions that illustrate how the deduced number of parents changes based on different numbers of distinct gametes or haplotypes observed within a brood.

3. Results

We developed three computer programs designed to estimate: (1) the number of parents that contributed to a half-sib progeny array and (2) the number of offspring that must be sampled from an array to detect at least one gamete from each parent. The results from each program are presented in turn, and potential applications of the programs are examined in light of empirical data. A flowchart illustrating the relationships between the programs is presented in Fig. 1.

(i) *BROOD*

Our program for determining necessary sample sizes is termed BROOD. We employed two statistics (n and n^*) to measure the effects of polymorphism on sampling regimes for parentage assessment of half-sib broods. These statistics differ in that n is with respect to alleles that are defined by a particular genetic marker, whereas n^* is with respect to quintessential alleles. Thus, n approaches n^* as the number of differentiable allelic states in the population approaches infinity. BROOD simulations illustrate the effects of genetic polymorphism and the number of unshared parents on these two statistics (Figs. 2–4).

As expected, $\bar{n}$ increased as the number of distinct alleles in the population increased (Fig. 2). The difference between $\bar{n}$ and $\bar{n}^*$ decreased as the level of marker polymorphism increased (Fig. 2). This is understandable because, in principle, $\bar{n}$ should approach $\bar{n}^*$ as allelic variation becomes large.

Uniform allele frequency distributions in the adult

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

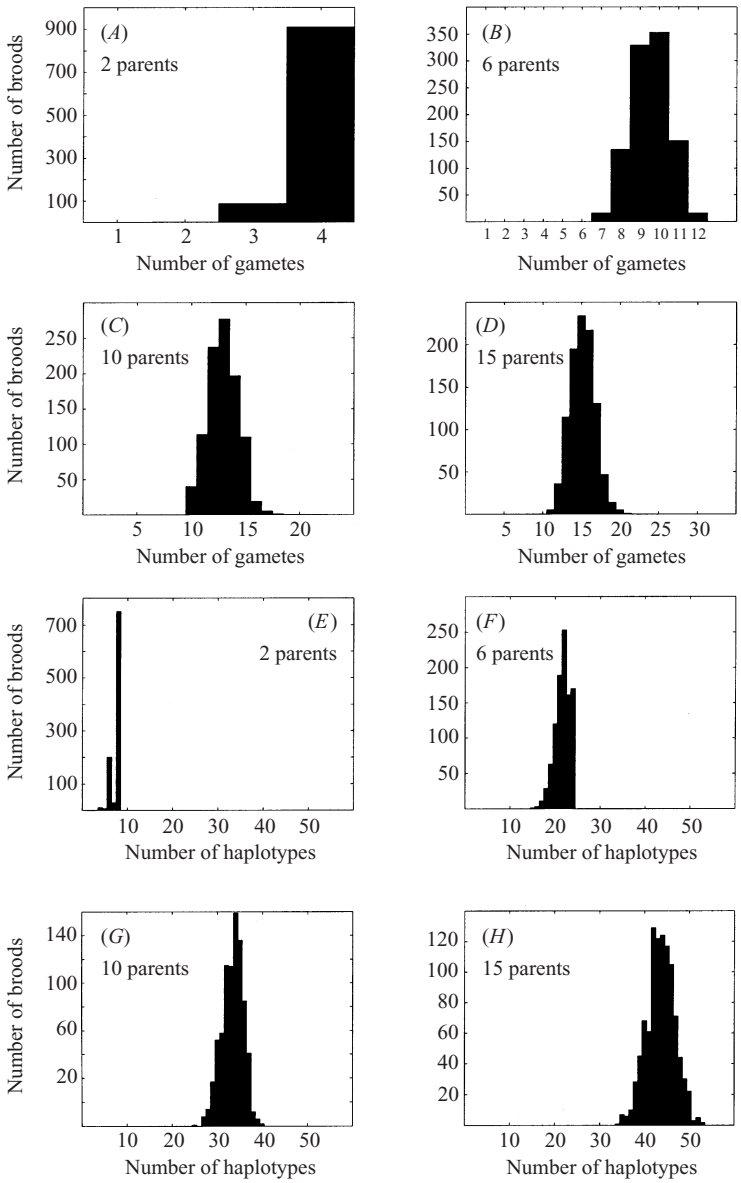

Fig. 5. Examples of graphical outputs from the GAMETE (*A–D*) and HAPLOTYPE (*E–H*) simulations. Shown are the number of outcomes (among 1000 trials) wherein the indicated number of marker-identified gametes or haplotypes in a brood resulted from varying numbers of unshared parents. Simulations were run under high-polymorphism conditions (see Table 2).

population did not entail substantially smaller $\bar{n}^*$ than skewed or highly skewed distributions (Fig. 2). This result, which may seem counterintuitive, arises because allele frequencies in a population have no bearing on Mendelian inheritance within a brood (e.g. if one parent of a clutch is heterozygous, about 50% of those offspring receive each of the two alleles regardless of allele frequencies in the adult population). However, skewed allele frequencies did decrease $\bar{n}$ slightly (Fig. 2). This results from the fact that under HWE, skewed as opposed to equitable allele frequencies give rise to more homozygous parents so that on average fewer alleles are found in each brood.

The second analysis (Fig. 3) supports the obvious notion that $\bar{n}$ and $\bar{n}^*$ (the number of offspring that must be sampled from a brood) increase as more unshared parents contribute to a half-sib brood. Notice that $\bar{n}$ asymptotes once the parental assemblage size is large enough to contain all allelic variation found in the population as a whole, whereas $\bar{n}^*$ increases without bound as the assemblage size increases. The simulations also support an intuitive notion that when the relative contributions of multiple unshared parents to a brood are skewed (as opposed to equitable; Keller & Reeve, 1994), more progeny must be tested to detect all parental gametes present. This sample size escalated dramatically if one or more of the unshared parents produced less than about 10% of the brood. For example, with four unshared parents, the upper bound on the 95% confidence

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

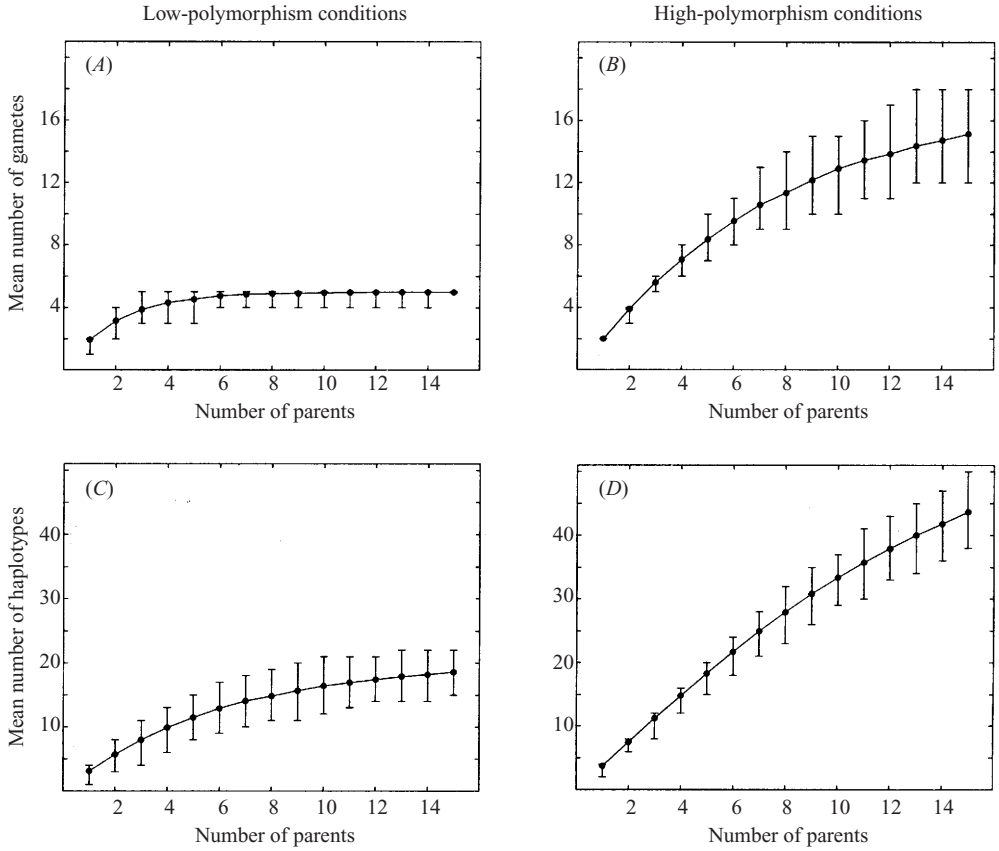

Fig. 6. Relationship between the number of unshared parents and the mean number of marker-identified gametes or haplotypes in a brood. Panels in the left and right columns were produced through simulations of low and high polymorphism conditions, respectively (Table 2). (*A*) and (*B*) were produced by GAMETES, whereas (*C*) and (*D*) were produced by HAPLOTYPE. Error bars are 95% confidence intervals. These plots were generated by MATLAB from an accumulation of the kinds of trials illustrated in Fig. 5.

interval for $\bar{n}$ increased from about 39 when all parents contributed equally to a brood to more than 92 when parental contributions were highly skewed.

Finally, all else being equal, $\bar{n}$ and $\bar{n}^*$ tended to increase as more loci or more alleles were monitored (Fig. 4). For example, with 25 alleles each at either two or four loci, $\bar{n}$ increased from about 51 to 61; and, with two loci each with either 10 or 25 alleles, $\bar{n}$ increased from about 43 to 51. Note, however, that $\bar{n}$ and $\bar{n}^*$ do not increase proportionately with the number of loci. For instance, with 25 alleles at each locus, $\bar{n}^*$ is about 66 at four loci and only 79 at 10 loci (Fig. 4). This disproportionate increase arises because genotypes were sampled from individuals, such that once a genotype is determined at the first locus, alleles at all other loci are sampled simultaneously.

(ii) *GAMETES and HAPLOTYPES*

The next two programs, termed GAMETES and HAPLOTYPES, will illustrate the concepts behind the simulations. In their current formulation, both programs assume an equal contribution of gametes from each unshared parent of a brood. Our empirical data on fish mating systems suggest that such uniform distributions may be reasonable as a first approximation, as for example when each of several females deposits a clutch of eggs in a male's nest. However, other distributions may be appropriate for different organisms (e.g. Harshman & Clark, 1998).

In addition to generating distributions of gametotype counts, each program also records the most likely number of unshared parents (parental assemblage size) for a given number of different gametes in a brood. GAMETES uses the single most informative locus from a suite of loci, whereas HAPLOTYPES incorporates multi-locus gametic data by permutating alleles across unlinked loci in the progeny array. Thus, for HAPLOTYPES the maximum possible number of distinct haplotypes from a single parent is 2^L, where L is the number of loci, and will always be greater than the number of gametes detected by GAMETES if $L > 1$. The maximum number of haplotypes in a progeny array is, thus, equal to the number of parents multiplied by 2^L. Hence, GAMETES can be viewed as a special case of HAPLOTYPE where $L = 1$.

One might expect that allele frequency distributions would play a large part in determining the estimated number of parents contributing to a brood. However,

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

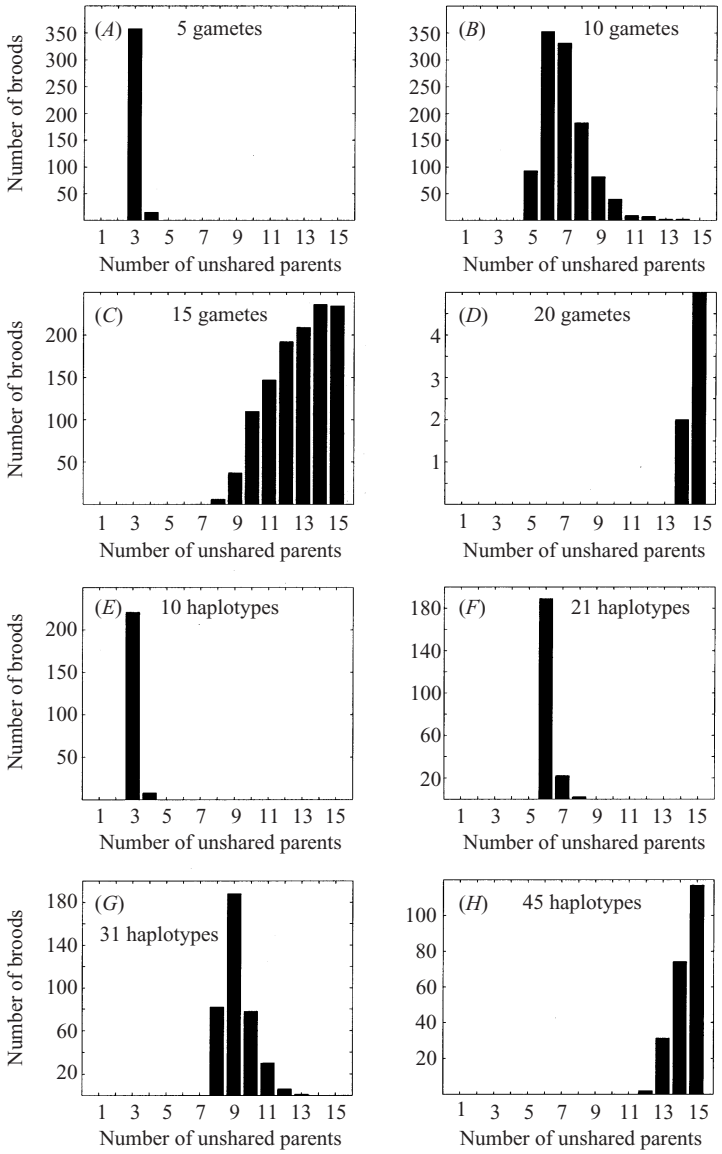

Fig. 7. Examples of additional graphical outputs from the GAMETES (*A–D*) and HAPLOTYPE (*E–H*) simulations. Shown are the number of outcomes wherein an indicated number of unshared parents contributed a given number of marker-identified gametes or haplotypes to a brood. Note that each distribution was created not as an inverse of a single assemblage distribution shown in Fig. 6, but as a grand total of all assemblage distributions ranging in size from 1 to 15. Thus, the size of the sample space varies (unlike Fig. 5).

in our simulations these estimates appear to be influenced more by the number of alleles than by the distributions *per se*. For example, nearly-normal allele frequency distributions similar to those found in the empirical studies of DeWoody *et al*. (1998) and Jones & Avise (1997*a*, *b*) gave similar results in the current analyses to those from the L-shaped empirical distributions reported by Luikart *et al*. (1998). Likewise, imprecise estimates of allele frequencies in the adult population did not strongly affect our estimates of the mean number of gametes or haplotypes in a progeny array (or the deduced numbers of parents), presumably because few parents have the rare alleles that would most often be missed in a population survey due to sampling error.

Examples of graphical results from the GAMETES and HAPLOTYPE simulations are shown in Figs. 5 and 7, and compilations of these respective classes of information are summarized in the corresponding Figs. 6 and 8. Fig. 7 shows the distributions of marker-identified gametes or haplotypes detected in broods with various numbers of parents and the population-genetic and sampling conditions specified (Table 2). From Fig. 8 it can be seen that gametic numbers (*A* and *B*) and haplotype numbers (*C* and *D*) in a brood both tended to increase with larger numbers of unshared parents. Furthermore, Fig. 8*D* shows, for example, that the confidence interval around the mean extended from 12–16 haplotypes when there were four unshared parents of a brood to 29–37 haplotypes

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

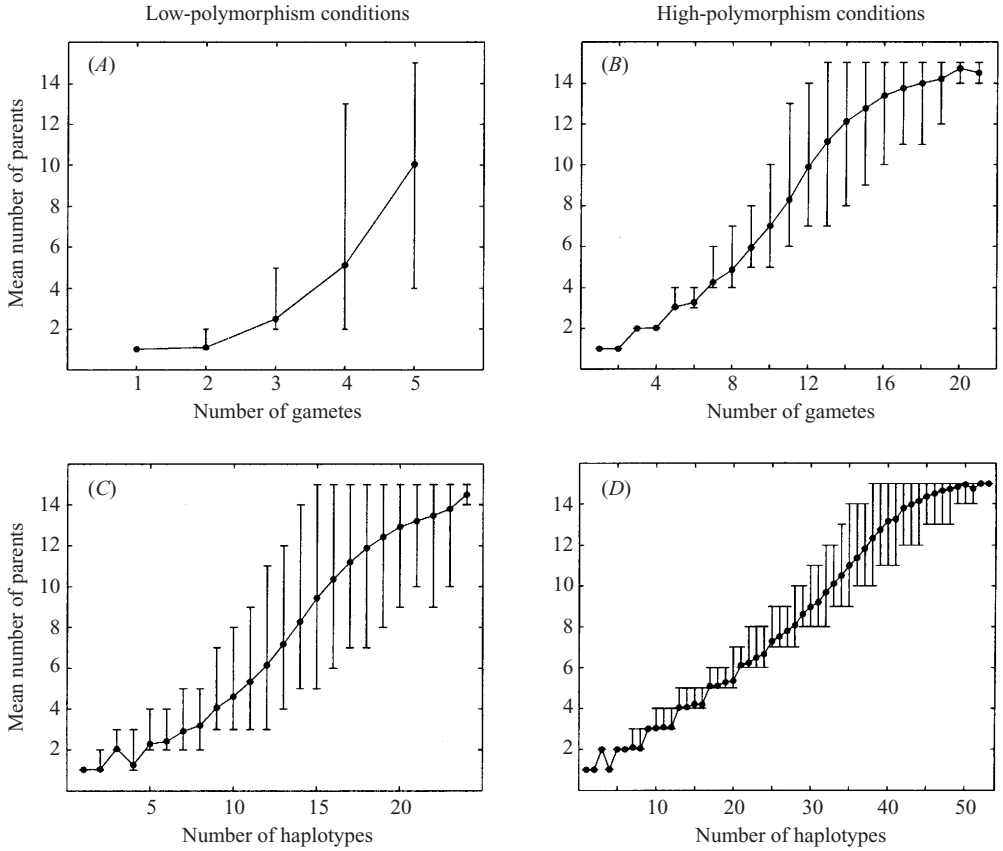

Fig. 8. Relationship between the mean number of marker-identified gametes or haplotypes in a brood and the number of unshared parents. Panels in the left and right columns represent conditions of low and high polymorphism, respectively (Table 2). (*A*) and (*B*) are from GAMETES, (*C*) and (*D*) from HAPLOTYPE. Error bars mark 95% confidence intervals around the mean number of unshared parents. These plots were generated by MATLAB from an accumulation of the kinds of trials illustrated in Fig. 7.

when there were 10 unshared parents. This means that 95% of the time, 10 unshared parents are expected to produce between 29 and 37 haplotypes.

In true empirical studies one is more likely to know the number of gametes contributed by unshared parents than to know the number of parents. Fig. 7 illustrates how the deduced number of parents can vary given different numbers of parental gametes or haplotypes detected in a brood. Except in cases where extremely few unshared parental gametes were present in a brood, the deduced number of parents spans a range of values. For example, Fig. 7*F* shows that under the conditions described, the presence of 21 haplotypes in a brood suggests six, seven or eight unshared parents, with relative likelihoods given by the peak heights. Thus, the most likely number of unshared parents is six, an outcome eight times more likely than the second most probable number (seven unshared parents). Fig. 8 compiles such information and shows the 95% confidence intervals.

Interestingly, one can also note from Fig. 8 that whereas GAMETES and HAPLOTYPES give similar results in terms of the mean number of unshared parents, the variance around the mean is much smaller when data from multiple loci (i.e. HAPLOTYPES) is utilized. Note too that the single-locus approach is particularly unstable under low-polymorphism conditions (Fig. 8*A*).

(iii) *Applications to real data*

To exemplify how these programs might assist in the design and interpretation of empirical research on parentage assessment of half-sib broods, empirical case studies will be presented for two fish species: the redbreast sunfish (*Lepomis auritus*; DeWoody *et al.*, 1998) and the tesselated darter (*Etheostoma olmstedii*; DeWoody *et al.*, unpublished data). In both species, individual males tend nests into which one or more females may lay eggs, and a primary question is how many females have contributed to a nest of embryos or fry whose father often is known. For both species, progeny cohorts and nest-attendant males were collected and assayed for microsatellite markers from each of multiple nests, and maternal gametes were deduced by subtraction. These two examples were chosen for illustration here because one represents a high-polymorphism and the other a low-

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

Table 3. *Microsatellite genetic loci from the darter* (Etheostoma olmstedii) *used in the empirical examples*

Locus designation	Number of alleles	Frequency of most common allele	Expected heterozygosity
EO4	10	0·27	0·820
EO6	2	0·54	0·499
EO12	4	0·88	0·224

polymorphism situation with respect to the microsatellite loci employed.

For the redbreast sunfish, two highly polymorphic loci (18 and 22 alleles each; frequencies described in DeWoody *et al*., 1998) were employed in the parentage assessments, and the number of embryos assayed per nest (25 nests) ranged from 10 to 175, with mean of 40. Using observed allele frequencies in the adult population, and assuming equal maternal contributions to a brood, BROOD simulations show that $\bar{n}$ is about 27 (upper 95% CI, 48) and $\bar{n}^*$ is 33 (upper 95% CI, 54). Thus, the empirical sample sizes originally employed indeed were sufficiently large (on average) to detect most if not all of the maternal alleles present within a particular nest under these conditions.

For the sunfish nests fathered by a single male, DeWoody *et al*. (1998) concluded from direct counts of deduced maternal gametes that minimally, between two and six mothers contributed to a brood. For example, one nest (LA12) had six different maternal alleles among 50 embryos sampled at one locus, meaning that no fewer than three females spawned in that nest. Single-locus simulations show that the 'adjusted' estimate of mothers for this nest was 3·8, with 95% confidence interval spanning three to six. Similarly, multi-locus simulations show that for another nest (LA28 phase B) with 16 di-locus haplotypes among 50 embryos, the adjusted number of mothers was 5·3 (95% CI, 4–8), whereas the minimum number based on direct genotypic count was four. Two points are evident from such examples. First, statistically adjusted numbers of deduced parents are larger than the minimal estimates from the direct-count method. Secondly, in these cases the adjusted estimates are close to the face-value estimates and do not alter biological conclusions appreciably.

In a continuing study of the tesselated darter, at the time of writing three microsatellite loci have been characterized, and polymorphism is relatively low in the adult population (Table 3). In this case, the darter broods have not yet been assayed genetically, so simulations based on the observed allele frequencies were conducted over a range of assemblage sizes to ascertain whether our sample sizes of offspring were appropriate. Results are shown in Fig. 9. For example, if we assume that two unshared parents contribute equally to each darter nest, we find that $\bar{n}$ is 7·3 and $\bar{n}^*$ is 9·0. Likewise, assuming seven unshared parents, these means are about 28 and 48, respectively. Thus, unless the number of unshared parents contributing to a half-sib darter nest is more than about seven, then samples of 50 offspring per nest should be adequate to capture parentage patterns in these fishes with the available markers.

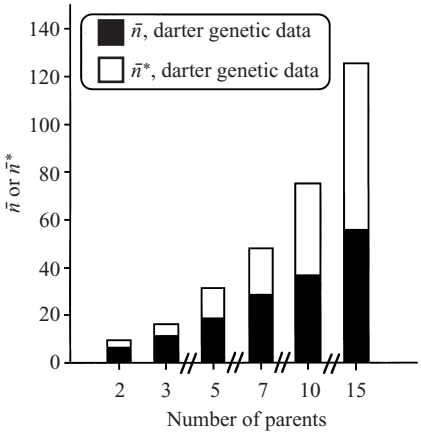

Fig. 9. Effects of the number of unshared darter parents on n and n^* when all other parameters were held constant. Empirically derived allele frequencies at the three darter loci are described in Table 3.

4. Discussion

For large broods consisting of mixtures of full-sibs and half-sibs, the simulations developed here permit appraisals of: (1) the mean number of offspring that must be sampled to detect all parental gametes in a brood and (2) the relationship between the number of distinct parental gametes in a brood and the true number of parents. Thus, simulations can assist in the design of empirical research on several aspects of genetic parentage in species with polygynous or polyandrous mating systems. These programs are available as MATLAB source code (and will soon be available as C code) on our website at www.genetics.uga.edu/popgen/parentage.html.

With regard to the first objective mentioned above, these simulations should aid in the design of sampling strategies when brood size is too large to permit feasible laboratory assay of all progeny in a clutch (as is often true for highly fecund species such as many amphibians, fishes and insects). With regard to the second objective, the simulation approach provides an improvement over conventional procedures of merely using allelic counts to estimate the number of unshared

https://doi.org/10.1017/S0016672399004000 Published online by Cambridge University Press

parents of a brood (e.g. DeWoody *et al*., 1998; Kellogg *et al*., 1998). For example, in a half-sib brood displaying 10 single-locus gametotypes, the minimum number of unshared parents by the direct-count method is five. However, in the simulated distributions, these 10 gametes more often truly arose from six, seven or eight unshared parents (Fig. 7*B*), and the 95% confidence interval spans 5–10 parents (Fig. 8*B*).

One major advantage to our approach is that potential unshared parents need not be sampled exhaustively because inferences are based largely upon the distribution of parental alleles in progeny arrays. Statistical approaches have been developed to resolve parentage issues in cases where genotypic data *are* exhaustive (i.e. Marshall *et al*., 1998), but our approach allows inferences in the absence of direct data on the unshared parents. Marshall *et al*. (1998) also correctly point out that typing errors are common in large-scale parentage studies; such typing errors may potentially bias estimates of the number of parents contributing to half-sib progeny arrays.

Explicit analytical approaches to maximum-likelihood (ML) offer another avenue for estimating numbers of parents contributing to a half-sib clutch. However, it has proved difficult to derive ML equations that take into account multiple unshared parents and multiple loci (Harshman & Clark, 1998). Current simulation methods are far less intensive computationally, and thus may be useful to biologists studying parentage in organisms with large clutch sizes.

We thank M. Asmussen, M. A. D. Goodisman, W. G. Hill, D. E. Pearse and D. Walker for comments on the manuscript. D. E. Fletcher and S. D. Wilkins collected the sunfish and darters described herein. Work was funded by the Pew Foundation and the University of Georgia.

References

Baker, R. J., Makova, K. D. & Chesser, R. K. (1999). Microsatellites indicate a high frequency of multiple paternity in *Apodemus* (Rodentia). *Molecular Ecology* **8**, 107–111.

Cobbs, G. (1997). Multiple insemination and male sexual selection in natural populations of *Drosophila pseudoobscura*. *American Naturalist* **111**, 641–656.

Coltman, D. W., Bowen, W. D. & Wright, J. M. (1998). Male mating success in an aquatically mating pinniped, the harbour seal (*Phoca vitulina*), assessed by microsatellite DNA markers. *Molecular Ecology* **7**, 627–638.

DeWoody, J. A., Fletcher, D. E., Wilkins, S. D., Nelson, W. S. & Avise, J. C. (1998). Molecular genetic dissection of spawning, parentage, and reproductive tactics in a population of redbreast sunfish, *Lepomis auritus*. *Evolution* **52**, 1802–1810.

Fitzsimmons, N. N. (1998). Single paternity of clutches and sperm storage in the promiscuous green turtle (*Chelonia mydas*). *Molecular Ecology* **7**, 575–584.

Griffiths, R. C., McKechnie, S. W. & McKenzie, J. A. (1982). Multiple mating and sperm displacement in a natural population of *Drosophila melanogaster*. *Theoretical and Applied Genetics* **62**, 89–96.

Harshman, L. G. & Clark, A. G. (1998). Inference of sperm competition from broods of field-caught *Drosophila*. *Evolution* **52**, 1334–1341.

Imhof, M., Harr, B., Brem, G. & Scholotterer, C. (1998). Multiple mating in wild *Drosophila melanogaster* revisited by microsatellite analysis. *Molecular Ecology* **7**, 915–918.

Jones, A. G. & Avise, J. C. (1997*a*). Microsatellite analysis of maternity and the mating system in the Gulf pipefish *Syngnathus scovelli*, a species with male pregnancy and sex role reversal. *Molecular Ecology* **6**, 202–213.

Jones, A. G. & Avise, J. C. (1997*b*). Polygynandry in the dusky pipefish *Syngnathus floridae* revealed by microsatellite DNA markers. *Evolution* **51**, 1611–1622.

Jones, A. G., Östlund-Nilsson, S. & Avise, J. C. (1998). A microsatellite assessment of sneaked fertilizations and egg thievery in the fifteenspine stickleback. *Evolution* **52**, 848–858.

Keller, L. & Reeve, H. K. (1994). Partitioning of reproduction in animal societies. *Trends in Ecology and Evolution* **9**, 98–103.

Kellogg, K. A., Markert, J. A., Stauffer, J. R. & Kocher, T. D. (1998). Intraspecific brood mixing and reduced polyandry in a maternal mouth-brooding cichlid. *Behavioral Ecology* **9**, 309–312.

Levine, L., Asmussen, M., Olvera, O., Powell, J. R., De La Rosa, M. E., Salceda, V. M., Gaso, M. I., Guzman, J. & Anderson, W. W. (1980). Population genetics of Mexican *Drosophila*. V. A high rate of multiple insemination in a natural population of *Drosophila pseudoobscura*. *American Naturalist* **116**, 493–503.

Luikart, G., Allendorf, F. W., Cornuet, J.-M. & Sherman, W. B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *Journal of Heredity* **89**, 238–247.

Marshall, T. C., Slate, J., Kruuk, L. E. B. & Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* **7**, 639–655.

Nason, J. D., Herre, E. A. & Hamrick, J. L. (1996). Paternity analysis of the breeding structure of strangler fig populations: evidence for substantial long-distance wasp dispersal. *Journal of Biogeography* **23**, 501–512.

Smouse, P. E. & Meagher, T. R. (1994). Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (Liliaceae). *Genetics* **136**, 313–322.

Taborsky, M. (1994). Sneakers, satellites, and helpers: parasitic and cooperative behavior in fish reproduction. *Advances in the Study of Behavior* **23**, 1–100.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer.