Title
Accuracy and Precision of Methods to Estimate the Number of Parents Contributing to a Half-Sib Progeny Array

Journal
Journal of Heredity, 92(2)

ISSN
0022-1503

Authors
Fiumera, AC
DeWoody, YD
DeWoody, JA
et al.

Publication Date
2001-03-01

DOI
10.1093/jhered/92.2.120

Accuracy and Precision of Methods to Estimate the Number of Parents Contributing to a Half-Sib Progeny Array

A. C. Fiumera, Y. D. DeWoody, J. A. DeWoody, M. A. Asmussen, and J. C. Avise

Molecular technologies have made feasible large-scale studies of genetic parentage in nature by permitting the genotypic examination of hundreds or thousands of progeny. One common goal of such studies is to estimate the true number of unshared parents who contributed to a large half-sib progeny array. Here we introduce computer programs designed to count the number of gametotypes contributed by unshared parents to each such progeny array, as well as assess the accuracy and precision of various estimators for the true number of unshared parents via computer simulation. These simulations indicate that under most biological conditions (1) a traditional approach (the multilocus MINIMUM METHOD) that merely counts the number of distinct haplotypes in offspring and divides by 2^L, where *L* is the number of loci assayed, often vastly underestimates the true number of unshared parents who contributed to a half-sib progeny array; (2) a recently developed HAPLOTYPES estimator is a considerable improvement over the MINIMUM METHOD when parental numbers are high; and (3) the accuracy and precision of the HAPLOTYPES estimator increase as marker polymorphism and sample size increase, or as reproductive skew and the number of parents contributing to the progeny array decrease. Generally, HAPLOTYPES-based estimates of parental numbers in large half-sib cohorts should improve the characterization of organismal reproductive strategies and mating systems from genetic data.

Molecular markers permit detailed investigations of biological parentage in nature (Avise 1994; Hughes 1998). Although numerous molecular studies have addressed paternity, maternity, and the genetic mating systems in avian and mammalian populations (see Andersson 1994; Birkhead and Møller 1998), far fewer have focused on fishes or other poikilothermic vertebrates (but see DeWoody et al. 1998; Jones et al. 1998; Kellogg et al. 1998; Parker and Kornfield 1996). Yet given the diverse spawning behaviors of fish (Taborsky 1994) and the potential to analyze data from the exceptionally large clutches they often produce, genetic studies of such groups should be highly informative about the broader evolution of organismal mating systems.

In many species, either males, females, or both sexes often mate with multiple partners. Such polygamous mating often may result in half-sib progeny arrays, where all offspring (e.g., within a nest) share one biological parent but not the other. For example, in many fish species, an attendant male builds and guards a nest into which multiple females deposit eggs which he then fertilizes (Taborsky 1994). Thus, barring cuckoldry, the offspring in such a nest comprise a collection of full-sib and half-sib progeny (i.e., a half-sib progeny array). In this progeny array, the attendant male is the shared parent, and the multiple females are the unshared parents.

Knowledge of the number of unshared parents contributing to a half-sib progeny array is vital to understanding behavioral, ecological, and other factors that may influence mating-system evolution. For example, do correlations exist between particular phenotypic characteristics (such as body size or coloration) and the number of successful mates? Does the number of surviving offspring in a clutch vary predictably as a function of a parent's success in obtaining mates? Such questions address the differential reproductive success of individuals, a central basis of sexual selection that should impact mating system evolution.

Fishes (as well as many other polygamous species with large clutches) present unprecedented opportunities for genetic analysis, but also some unique challenges

From the Department of Genetics, University of Georgia, Athens, GA 30602. J. A. DeWoody is currently at the Department of Forestry and Natural Resources, Purdue University, West Lafayette, Indiana. Y. D. DeWoody is currently at Purdue University, West Lafayette, Indiana. This work was supported by a National Institutes of Health training grant (to A.C.F.), by funds from the University of Georgia, by the Pew Foundation (to J.C.A.), and by the National Science Foundation (DEB-9906462) (to M.A.A.). Useful comments on the manuscript were provided by Mark Mackiewicz, Beth McCoy, Patty Parker, Devon Pearse, Brady Porter, and DeEtte Walker. Tigerin Peare, Jennifer Bollmer, and Patty Parker kindly provided unpublished turtle data. Address correspondence to Anthony Fiumera at the address above or e-mail: fiumera@arches.uga.edu. This paper was delivered at a symposium entitled "DNA-Based Profiling of Mating Systems and Reproductive Behaviors in Poikilothermic Vertebrates" sponsored by the American Genetic Association at Yale University, New Haven, CT, USA, June 17–20, 2000.

 92:120–126

seldom encountered in genetic parentage studies of mammals and birds. For example, a fish nest often may contain many thousands of offspring from multiple spawning events, with unknown parents (of at least one gender) drawn from a large adult population. Furthermore, it is seldom feasible to assay all of the progeny from each nest, so statistical issues inevitably arise concerning optimal sampling design in the genetic appraisals of parentage.

To begin to address these challenges, a package of computer simulation programs termed REPRODUCTIVES (composed of BROOD, GAMETES, and HAPLOTYPES) was developed to estimate, from codominant molecular markers, the true number of unshared parents contributing to a half-sib progeny array (DeWoody et al. 2000a). BROOD is designed to determine the average sample sizes of progeny necessary to detect the gametic contribution of each unshared parent. Brood reports two values, $\bar{n}$ and $\bar{n}^*$, as well as the 95% confidence limits of these values. The average number of offspring that must be sampled to detect all marker-specific gametes contributed by the unshared parents is termed $\bar{n}$, whereas $\bar{n}^*$ is the average number of offspring that need to be sampled to detect all true gametes contributed by the unshared parents (assuming that all gametes can be differentiated). GAMETES and HAPLOTYPES are single- and multilocus estimators, respectively, of the number of unshared parents of such an array. Several other methods have been designed to assess parentage using genetic marker data (Griffiths et al. 1982; Harshman and Clark 1998; Kellogg et al. 1998; Levine et al. 1980; Marshall et al. 1998; Parker and Kornfield 1996). Two of these, the traditional single- and multilocus MINIMUM METHOD, are simple to implement and have been applied frequently in empirical studies (e.g., DeWoody et al. 1998; Kellogg et al. 1998; Parker and Kornfield 1996).

To draw sound biological conclusions from empirical genetic data, the statistical methods employed must also be sound. Any useful estimator of parental numbers, for example, should be both accurate and precise. Thus the goals of this study are to introduce computer programs to count the number of distinct gametotypes contributed by unshared parents to a half-sib progeny array and to compare the accuracy and precision of alternative procedures for estimating the true number of unshared parents from such gametotypic counts.

Materials and Methods

Counting Gametotypic Numbers

We have developed three computer programs (COUNTS LOW, COUNTS MEDIUM, and COUNTS HIGH) to tally the number of genetically distinct gametotypes (single-locus gametes or multilocus haplotypes) contributed by unshared parents to a half-sib progeny array. Each program assumes that the diploid genotypes of one parent (the shared parent) and of varying numbers of its progeny are known. This is often the case in the empirical literature where, for example, molecular assays have been used to genotype a parental guardian and some of the progeny from his or her nest. The COUNTS programs are available as MATLAB source code at www.genetics.uga.edu/popgen/parentage.html.

For each offspring in a nest, the alleles inherited from the unshared parent are deduced by subtraction [under the rationale that, barring de novo mutation, any allele in the progeny not displayed by the known (shared) parent must have come from the unshared parent]. This subtraction procedure always yields an unambiguous allelic assignment for the unshared parent, except when an offspring displays the same heterozygous genotype as its shared parent. [In a random mating population at Hardy–Weinberg equilibrium for n alleles, this occurs with probability $\Sigma_{i=1}^{n} p_i^2(1 - p_i)$, where p_i is the frequency of the A_i allele in the population (Fiumera AC and Asmussen MA, submitted).]

In these ambiguous cases, each COUNTS program utilizes different approaches to tally the number of distinct gametotypes. For simplicity, consider the case where only a single locus has been assayed; similar procedures are followed if the offspring and parent have been genotyped at multiple loci. COUNTS LOW (the most conservative method) treats this ambiguous locus as missing data; COUNTS MEDIUM randomly assigns one of the two possible alleles as the presumed contribution from the unshared parent and then tallies this contribution if it has not been detected in other offspring sampled from the nest; and COUNTS HIGH randomly assigns to the unshared parent whichever of the two possible alleles had not yet been previously detected in other offspring.

For example, suppose that the known parent and one of its progeny are both A_1A_2 heterozygotes at a given locus. COUNTS LOW ignores the genotype of this offspring at this locus, while COUNTS MEDIUM randomly assigns either the A_1 or the A_2 allele (with probability 0.5 for each) as the contribution of the unshared parent and tallies a new distinct gametotype only if it had not been previously detected. COUNTS HIGH assigns whichever of these two alleles had not been sampled previously in other offspring. If neither allele had been sampled, COUNTS HIGH randomly assigns either A_1 or A_2. When no offspring are identically heterozygous to their known shared parent, all three COUNTS programs report the same value. Each of the COUNTS programs reports the number of multilocus gametotypes (i.e., haplotypes) across all loci and the number of single-locus gametotypes (i.e., alleles) from only the most informative locus (i.e., the single locus with the largest number of alleles attributable to the unshared parents).

Description of Estimators

Four statistical estimators of the numbers of unshared parents contributing to a half-sib progeny array are considered in this report; GAMETES and HAPLOTYPES as well as the single- and multilocus MINIMUM METHOD. GAMETES and HAPLOTYPES (see DeWoody et al. 2000a for details) are computer simulations that yield single- and multilocus estimators, respectively, of these parental numbers given genotypic data from a shared parent and varying numbers of his or her offspring. Both programs generate adult "breeding" populations based on specified allele frequencies and the corresponding Hardy–Weinberg genotypic frequencies. A single shared parent and from one to N unshared parents (N is defined by the user) are then chosen at random to be the parents of the half-sib progeny in a nest. Each nest of specified size is generated assuming Mendelian inheritance and equal offspring contributions by each unshared parent. The progeny array is then sampled (according to a sample size specified by the user) and the number of different gametotypes counted. This process is repeated thousands of times to generate frequency distributions for the number of distinct gametotypes contributed by N unshared parents given the designated parameters. Finally, the number of distinct gametotypes attributed to unshared parents in an actual empirical dataset (tallied using the COUNTS programs) is compared to the

Table 1. Parameters and conditions utilized in the computer programs BROOD, GAMETES, HAPLOTYPES, and in the current simulations to calculate accuracy and precision in the estimators of the number of unshared parents contributing to a half-sib progeny array

BROOD simulations	
Number of loci	2 and 4[a]
Number of alleles per locus	5, 10, 15, 25
Allele frequencies	Equal[b]
Number of parents	5
Parental contribution	Skewed 60/10/10/10/10 (i.e., 60% for 1 parent, 10% each for remaining 4 parents)
Accuracy and precision simulations[c]	
Number of parents	1 to 10
Parental contribution (skewed contr.)	Equal and skewed
2 parents	75/25
3 parents	70/15/15
4 parents	70/10/10/10
5 parents	60/10/10/10/10
6 parents	50/10/10/10/10/10
7 parents	40/10/10/10/10/10/10
8 parents	40/(8.57 each for 7 remaining parents)
9 parents	30/(8.75 each for 8 remaining parents)
10 parents	30/(7.77 each for 9 remaining parents)
Sample size[d] $<\bar{n}$, $\bar{n}$, $\bar{n}^*$, $>\bar{n}^*$	
GAMETES and HAPLOTYPES simulations[c]	
Number of parents	1 to 15
Parental contribution	Equal
Sample size[d]	$<\bar{n}$, $\bar{n}$, $\bar{n}^*$, $>\bar{n}^*$

[a] One example was completed with four loci, each having five alleles at equal frequencies.

[b] Two examples were conducted using empirically determined allele frequencies (see Table 2).

[c] Number of loci, number of alleles and allele frequencies were identical to those described for the BROOD simulations.

[d] $\bar{n}$ is the average number of offspring that must be sampled to detect all marker specific gametes contributed by the unshared parents; $\bar{n}^*$ is the average number of offspring that need to be sampled to detect all *true* gametes contributed by the unshared parents. Sample sizes of $<\bar{n}$ (1 standard deviation below $\bar{n}$), $\bar{n}$, $\bar{n}^*$ and $>\bar{n}^*$ (upper 95% confidence limit of $\bar{n}^*$) were obtained from the BROOD simulations. These correspond to samples sizes of approximately 25, 50, 60 and 100 offspring, respectively.

simulated distributions to determine the most likely number of unshared parents (reported as the mode of the distribution) who contributed to a particular half-sib progeny array.

For comparison, the performance of two additional estimators (the traditional single-locus and the multilocus MINIMUM METHOD; e.g., Parker and Kornfield 1996; Kellogg et al. 1998, respectively) were also examined. The single-locus MINIMUM METHOD estimates the suspected number of unshared parents for a half-sib progeny array as the total number of different single-locus gametotypes (i.e., alleles) attributed to the unshared parents, divided by two. This number is rounded up if necessary. Under this method, as traditionally applied, only the data from the most informative genetic locus are used. The multilocus MINIMUM METHOD estimates the suspected number of unshared parents as the total number of different multilocus gametotypes (i.e., haplotypes) attributed to the unshared parents, divided by 2^L, where L is the number of loci assayed. Again, this number is rounded up if necessary. The number of gametotypes for both the single-locus and multilocus MINIMUM METHOD was tallied using COUNTS LOW.

Table 2. Genetic data for the two empirical case studies described in the text

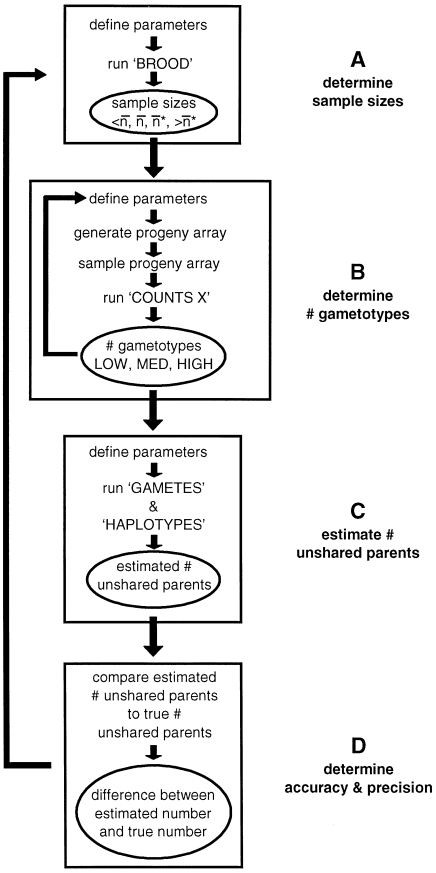
	Number of alleles	Allele frequencies	Effective number of alleles[a]	Reference
Redbreast sunfish (high polymorphism)				
Locus 1 (*RB7*)	22	.180 .133 .086 .070 .062 .055 .055 .055 .047 .039 .039 .031 .031 .023 .023 .023 .008 .008 .008 .008 .008 .008	11.8	DeWoody et al. 1998
Locus 2 (*RB20*)	18	.172 .140 .117 .117 .086 .078 .070 .047 .047 .031 .023 .016 .016 .008 .008 .008 .008 .008	9.8	
Sum	40		21.6	
Green turtle (low polymorphism)				
Locus 1 (*CC117*)	12	.206 .167 .103 .103 .103 .103 .051 .051 .039 .026 .026 .026	8.2	Peare T, et al., unpublished data
Locus 2 (*CM3*)	8	.385 .192 .192 .103 .051 .026 .026 .026	4.2	
Sum	20		12.4	

[a] Calculated using equation 2.41 in Hedrick (1985).

Estimates of Accuracy and Precision

Computer simulations were used to evaluate the accuracy and precision of the single- and multilocus MINIMUM METHOD (using COUNTS LOW), and of GAMETES and HAPLOTYPES (using COUNTS LOW, COUNTS MEDIUM, and COUNTS HIGH) estimates for the number of unshared parents that contributed to each half-sib array. Henceforth these outcomes will be referred to as HAP/LOW (for estimates from HAPLOTYPES using COUNTS LOW), and so forth. The effects of the number of loci, number of alleles per locus, number of unshared parents contributing to a progeny array, reproductive skew among parents, and the empirical sample sizes of progeny (see Table 1) were all investigated for their effects on both the accuracy and precision of each method. In addition to the hypothetical allele frequency distributions described in Table 1, the actual allele frequencies were employed from two empirical case studies (Table 2): one involving a population of redbreast sunfish with relatively high polymorphism (DeWoody et al. 1998), and the other involving green turtles with lower genetic variation (Peare T, et al., unpublished data).

Briefly, for each simulation 50 replicate progeny arrays were generated under a set of conditions defined in Table 1, knowing the user-defined (true) number of unshared parents. Each COUNTS program then tallied the number of distinct gametotypes in each progeny array, from which the number of unshared parents was estimated using GAMETES, HAPLOTYPES, and the single- and multilocus MINIMUM METHOD. The accuracy of each of these methods was calculated as the mean difference (in the 50 replicates) between the estimated and the true number of parents. Precision was estimated as the sample variance (across the 50 replicates) in the difference between the estimated and the true number of parents. Figure 1 provides a flowchart summarizing the procedures.



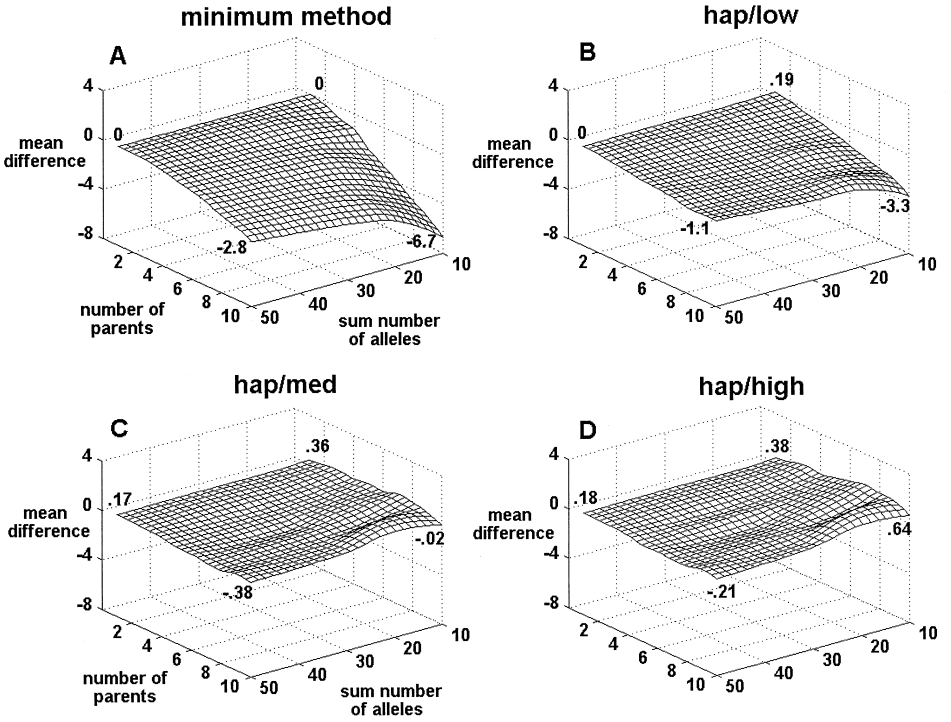


Figure 2. Effects of varying levels of marker polymorphism and numbers of unshared parents on the accuracy of the multilocus **(A)** MINIMUM METHOD, **(B)** HAP/LOW, **(C)** HAP/MEDIUM, and **(D)** HAP/HIGH procedures for estimating the number of unshared parents contributing to a half-sib progeny array. For these surfaces, simulations involving all four sample sizes and from both equal and skewed parental contributions (Table 1) were combined. Numbers inside each graph represent values at the four corners.

Figure 1. Flowchart of steps taken to determine the accuracy and precision of various estimators of parental contributors to a half-sib progeny array. Each box depicts a distinct procedure and the circled text indicates the results obtained and used in the subsequent procedure. Procedures A–D were completed for each set of parameters defined in Table 1. **(A)** The number of offspring to be sampled from each progeny array was determined using BROOD simulations. **(B)** Simulated progeny arrays were generated and the specified numbers of progeny (from procedure A) were sampled. The number of gametotypes observed in the progeny sample was then tallied using COUNTS LOW, MEDIUM, and HIGH. Procedure B was repeated 50 times, assuming 1–10 unshared parents and both equal and skewed parental contributions for each defined set of parameters (Table 1). **(C)** From the number of gametotypes counted in B, the number of unshared parents contributing to each progeny array was estimated using GAMETES, HAPLOTYPES, or the single- or multilocus MINIMUM METHOD as described in the text. **(D)** To determine the accuracy and precision of these various methods, the estimated number of unshared parents (from C) then were compared to the true number of parents as defined by the simulations for each of the different statistical estimators.

Results

Overall Accuracy of the Estimators

As expected, the multilocus estimator (HAPLOTYPES) was more accurate than the single-locus estimator (GAMETES) under virtually all conditions. In addition, the multilocus MINIMUM METHOD outperformed the single-locus MINIMUM METHOD, but only by a small margin. Therefore further analyses will focus on HAPLOTYPES versus the traditional multilocus MINIMUM METHOD. The overall accuracy and precision of HAPLOTYPES and the MINIMUM METHOD are presented in Figures 2–5.

Both the accuracy and the precision of all estimators increase as either the polymorphism of the markers (i.e., number of equally frequent alleles) increases or the number of unshared parents contributing to the progeny array decreases. The MINIMUM METHOD is highly accurate when three or fewer unshared parents contribute to a half-sib progeny array, but increasingly underestimates the true number of unshared parents for nests with larger parental numbers (Figure 2A). Though often an underestimate, the MINIMUM METHOD does tend to have a low sample variance (i.e., it is precise; Table 3). By contrast, HAPLOTYPES (using any of the COUNTS programs) is more accurate over a wider range of parameters than the MINIMUM METHOD (Figure 2). In many cases its performance is remarkably good. For example, using two loci each with 25 alleles, HAP/LOW is only 1.1 parents away from the true number, on average, when there are 10 unshared parents and progeny sample sizes are high (Figure 3B).

Overall, HAP/LOW, HAP/MED, and HAP/HIGH appear fairly similar in their accuracy over much of the parameter space considered (Figure 2B–D). HAP/LOW performs best when fewer than five unshared parents contributed to a nest, whereas HAP/MED and HAP/HIGH perform best when more than five unshared parents were involved (Figure 2B–D). However, HAP/MED tends to be an overestimate when only a few parents have contributed to a progeny array and HAP/HIGH tends to be an overestimate over most of the biological conditions investigated. HAP/MED does slightly outperform HAP/LOW, on average, when there are many parents and genetic polymorphism is low, but there is very low precision in this area of the parameter space (Figure 3C,D). In addition, HAP/LOW has the appeal of taking a simple and conservative stance by acknowledging the lack of information provided by offspring with the same heterozygous genotype as the shared parent, as compared to HAP/MED which randomly assigns one of the possible gametotypes. Given the inherent difficulty of estimating a large number of unshared parents accurately and precisely when marker polymorphism is low, HAP/LOW is probably the most appropriate estimator for most biological situations. Therefore the remainder of the analyses will focus on the use of HAP/LOW.

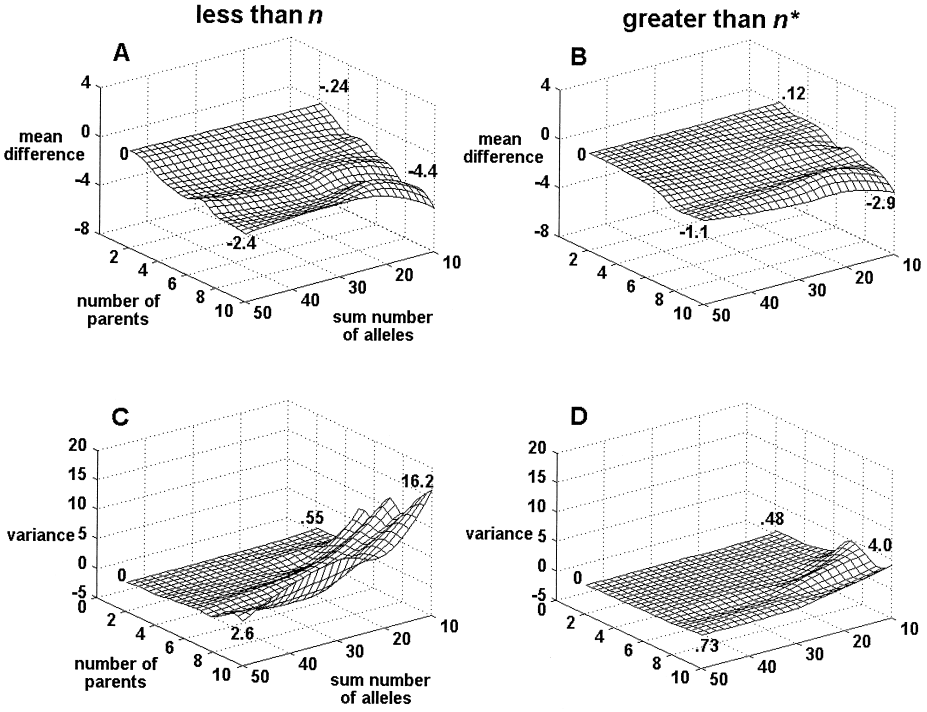


Figure 3. Effects of progeny sample size on the accuracy **(A,B)** and precision **(C,D)** of HAP/LOW under varying levels of marker polymorphism and numbers of unshared parents, given that parental contributions were skewed. Numbers inside each graph represent values at the four corners. Figures A and C correspond to a progeny sample size of less than $\bar{n}$ (approximately 25 in this example); B and D correspond to a sample size of greater than $\bar{n}^*$ (approximately 100 in this example).

Skew in Parental Contributions and Sample Size

HAPLOTYPES assumes that all unshared parents have contributed equally to a nest, so it is important to know how well this estimator performs when relative parental contributions are unequal. As expected, HAP/LOW tends to underestimate the true number of parents by a larger magnitude when parental contributions are skewed as compared to uniform (Figure 4). This effect can be partially mitigated by sampling more offspring from each nest (although the 95% confidence limits no longer capture the true value 95% of the time). As the number of offspring sampled increased from less than $\bar{n}$ to greater than $\bar{n}^*$ (see of Table 1 for definitions), the accuracy of HAP/LOW improved greatly (Figure 3A,B). Even more dramatic was the effect on the precision of the estimate. As sample size increased, the variance in the error estimate of HAP/LOW decreased (Figure 3C,D), sometimes dramatically.

To emphasize the importance of precision in an estimator, and how sample size affects the precision of HAP/LOW, consider one particular example. For the case of two loci each with 15 alleles at equal frequencies, and with 10 unshared parents contributing to each progeny array, progeny sample sizes of either less than $\bar{n}$, or of greater than $\bar{n}^*$, both underestimated the true number of unshared parents by approximately one individual, on average. However, the precision of the estimate was much higher with larger sample sizes. When a sample of less than $\bar{n}$ offspring (30 per nest in this particular example) was analyzed, HAP/LOW never estimated the true number of parents perfectly; it came within one parent of the true number 50% of the time; and it erred by more than three parents in 38% of the simulations (Figure 5A). However, for sample sizes of greater than $\bar{n}^*$ offspring (104 in this example), HAP/LOW perfectly estimated the true number of parents in 22% of the simulations; was within one of the true number in 62% of cases; and never erred by more than three unshared parents (Figure 5B).

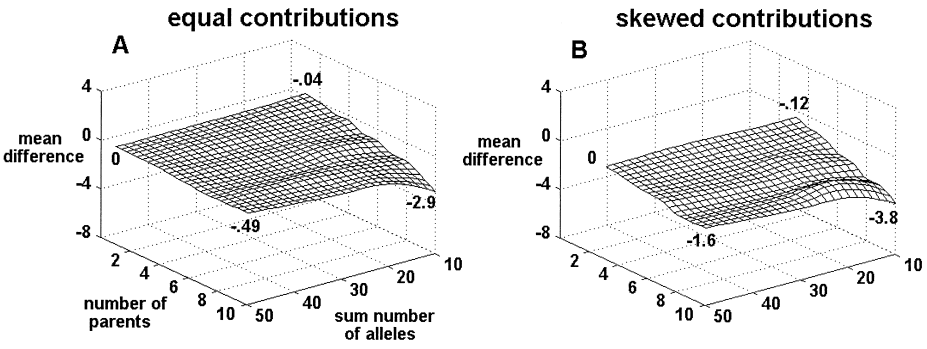


Figure 4. Effects of skew in parental contributions on the accuracy of HAP/LOW under varying levels of marker polymorphism and numbers of unshared parents. For these surfaces, simulations including all four sample sizes were combined. Numbers inside the graphs represent values at the corners.

Additional Loci and Empirical Examples

Simulations were also conducted using the allele frequencies from two empirical datasets (Table 2). In the case of the redbreast sunfish, there were 21.6 effective alleles (calculated as $\hat{n}_e = 1/\Sigma_{i=1}^{n} \hat{p}_i^2$ from equation 2.41 in Hedrick 1985) summed across the two loci assayed (Table 2). The simulations for these real data were consistent with those involving hypothetical data that arbitrarily assumed two loci each with 10 equally frequent alleles (i.e., 20 effective alleles total). For example, using HAP/LOW, the mean differences between the estimated and the true number of unshared parents were -0.8 and -0.9 for the real and hypothetical allele frequency distributions, respectively (Table 3). In addition, estimates from a green turtle population in which the empirical effective number of alleles was 12.4 were comparable to the case with two loci each with five equally frequent alleles (i.e., 10 effective alleles total) (Table 3).

Thus the effective number of alleles might appear to be a useful yardstick for the anticipated accuracy and precision of HAP/LOW across different allele frequency distributions. For example, assuming that all alleles at a given locus were equally frequent, two loci each with 10 alleles (20 effective alleles) provided more accurate and precise results than two loci each with 5 alleles (10 effective alleles). However, the effective number of alleles by itself did not invariably predict the simulation outcomes. For example, two loci each with 10 alleles provided much more accurate and precise results than did four loci each with 5 alleles (Table 3), despite

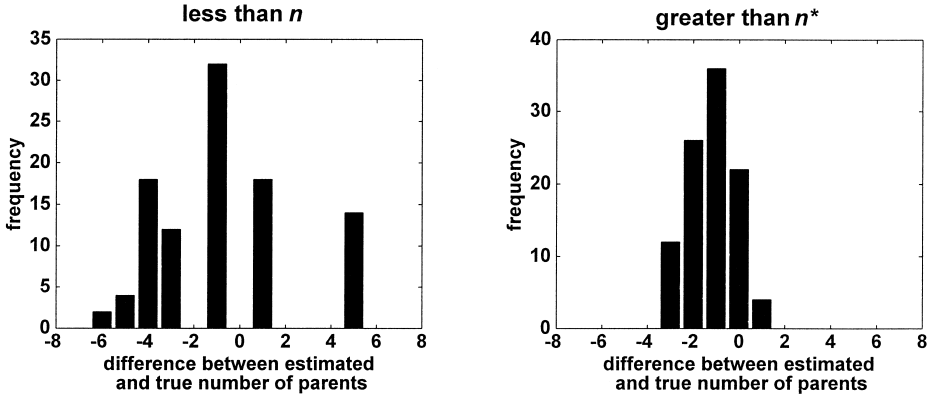


Figure 5. Example of the effects of progeny sample size on the precision of HAP/LOW. The frequency distribution in the difference between the estimated and the true number of unshared parents is shown for (left panel) sample sizes of less than $\bar{n}$ (30 in this example), and (right panel) sample sizes at the upper 95% confidence interval of $\bar{n}^*$ (104 in this example). The mean difference between the estimated and true number of unshared parents is approximately −1.0 for both distributions. This example was generated using two loci each with 15 equally frequent alleles, and assuming that 10 unshared parents contributed equally to each half-sib progeny array.

the fact the effective number of alleles was 20 in both cases.

Discussion

Using molecular markers such as microsatellites, it is now feasible to genotype hundreds or even thousands of individuals in an empirical survey of genetic parentage. Such technical capabilities open novel opportunities (as well as statistical challenges) for examining fishes or other species with high fecundities, where hundreds or thousands of embryos may be present within the nest of a single attendant parent. For instance, estimating the true number of parents contributing to a half-sib progeny array often can provide a far more informative picture of reproductive behaviors and of the mating system than can traditional genetic analyses that simply exclude particular adults as potential parents of an offspring array.

Here we have evaluated the effectiveness of alternative methods to estimate the number of unshared parents who contributed to a large half-sib progeny array. Computer simulations were used to determine the accuracy and precision of these various estimators under multiple biological scenarios. Parameters used in the computer simulations included the level of marker polymorphism, the sample size of progeny, the number of unshared parents, and the numerical skew in parental contributions to the pool of progeny within a nest.

Under virtually all conditions investigated, the multilocus estimator HAPLOTYPES proved to be more accurate than the multilocus MINIMUM METHOD, or either of the single-locus estimators (GAMETES or the MINIMUM METHOD). The traditional single- and multilocus MINIMUM METHOD were highly accurate when few parents (typically three or fewer) contribute to a half-sib progeny array; otherwise they could greatly underestimate true parental numbers. A low variance of the MINIMUM METHOD does make it attractive for cases where the investigator wishes merely to determine if a progeny array consists entirely of full sibs or, alternatively, if it includes half-sib progeny as well. However, a maximum-likelihood method already exists to assess the relative likelihood of single versus multiple paternity (Kichler et al. 1999).

Most of the above conclusions stem from computer simulations involving two marker loci, each with varying numbers of equally frequent alleles. Other conditions normally may have greater biological realism. Fortunately the effective number of alleles in particular datasets often (but not invariably) provided a useful predictor of relative accuracy and precision in the statistical estimators of the true parental numbers. In general, the use of multiple loci improved the power of these estimation methods. However, even with equal total numbers of effective alleles, a few highly polymorphic loci could yield more accurate and precise estimates of parental numbers than could many loci with low polymorphism.

It is important to recognize the limitations of any of these statistical methods. If the true number of parents contributing to a progeny array is large (greater than about five unshared parents), one cannot expect accurate and precise estimates of parental numbers from limited progeny samples, or from markers with low polymorphism. Thus in practice, serious attempts should be made to assay highly polymorphic markers and to sample at least $\bar{n}$ progeny (but preferably either $\bar{n}^*$ or the upper 95% confidence limit of $\bar{n}^*$ from BROOD). Given this caveat, the REPRODUCTIVES package appears to offer reasonably accurate and precise estimates of parental numbers over a wide range of biological conditions. Particularly for large half-sib clutches that are the product of many unshared parents, use of this computer program should substantially improve estimates of the number of parents contributing to a nest. In fact, an empirical "ground truthing" of HAPLOTYPES demonstrated exact agreement between the estimated number of unshared parents from a sample of 20 offspring and the "true" number determined by virtually exhaustive sampling of 906 progeny from a single fish nest (DeWoody et al. 2000b). This result held even though one of the three unshared parents contributed nearly

Table 3. Summary of the accuracy (mean difference) and precision (variance) obtained by the HAPLOTYPES and multilocus MINIMUM METHOD approaches for estimating true parental number in a half-sib progeny array using COUNTS LOW to tally the number of distinct gametotypes

Dataset	Σ effective number of alleles	Estimator	Mean difference	Variance
2 loci (all data)	10–50	HAPLOTYPES	−0.9	2.1
		MINIMUM	−1.9	0.3
Turtles[a]	12.4	HAPLOTYPES	−1.2	3.6
		MINIMUM	−2.6	0.4
2 loci (5 alleles each)	10	HAPLOTYPES	−1.4	4.4
		MINIMUM	−3.1	0.3
4 loci (5 alleles each)	20	HAPLOTYPES	−2.1	2.7
		MINIMUM	−3.8	0.1
2 loci (10 alleles each)	20	HAPLOTYPES	−0.9	1.8
		MINIMUM	−1.8	0.4
Sunfish[b]	21.6	HAPLOTYPES	−0.8	1.9
		MINIMUM	−1.7	0.4

The simulations here involved varying the total effective number of alleles (summed across loci).

[a] From Peare T, et al. (unpublished data).

[b] From DeWoody et al. (1998).

50% of the offspring to that nest. More generally the surfaces in Figures 2–4 provide information on parameter spaces where any empirically based estimates of parental numbers in half-sib clutches are (and are not) secure.

Several other statistical approaches have been developed to assess parentage using data from molecular markers (e.g., Griffiths et al. 1982; Harshman and Clark 1998; Kellogg et al. 1998; Levine et al. 1980; Marshall et al. 1998; Parker and Kornfield 1996), and some have overlapping goals with the current approach. For example, the CERVUS program of Marshall et al. (1998) is a maximum-likelihood approach that can be utilized to estimate the true number of unshared parents contributing to a half-sib progeny array, but it requires extensive knowledge (beyond what is normally available in parentage studies of many organisms in nature) on the genotypes of the potential parents. The analytical approach of Levine et al. (1980) assumes that all parental alleles are observed in the sample of the progeny and therefore does not account for the effects of sampling varying numbers of offspring. The Harshman and Clark (1998) approach estimates remating frequency and sperm precedence by assuming a geometric decline in fertilization success by successive unshared males, but this assumption may not be realistic for many organisms.

The REPRODUCTIVES package (DeWoody et al. 2000a), considered here, differs from previous methods in that it (1) corrects for the possibility that parents share alleles, (2) accounts for empirical sample sizes from a progeny array, and (3) only requires knowledge of the allele frequencies in the adult population (i.e., the unshared parents contributing to a progeny array need not be actually sampled). Like other focused statistical methods, when applied in appropriate biological settings, this package should facilitate the use of molecular data in addressing a variety of questions involving genetic parentage, sexual selection, and alternative reproductive strategies in a wide range of creatures with large, half-sib clutches.

References

Andersson M, 1994. Sexual selection. Princeton, NJ: Princeton University Press.

Avise JC, 1994. Molecular markers, natural history and evolution. New York: Chapman & Hall.

Birkhead TR and Møller AP, 1998. Sperm competition and sexual selection. New York: Academic Press.

DeWoody JA, DeWoody YD, Fiumera AC, and Avise JC, 2000a. On the number of reproductives contributing to a half-sib progeny array. Genet Res 75:95–105.

DeWoody JA, Fletcher DE, Wilkins SD, Nelson WS, and Avise JC, 1998. Molecular genetic dissection of spawning, parentage, and reproductive tactics in a population of redbreast sunfish, *Lepomis auritus.* Evolution 52: 1802–1810.

DeWoody JA, Walker D, and Avise JC, 2000b. Genetic parentage in large half-sib clutches: theoretical estimates and empirical appraisals. Genetics 154:1909–1912.

Griffiths RC, McKechnie SW, and McKenzie JA, 1982. Multiple mating and sperm displacement in a natural population of *Drosophila melanogaster.* Theor Appl Genet 62:89–96.

Harshman LG and Clark AG, 1998. Inference of sperm competition from broods of field-caught *Drosophila.* Evolution 52:1334–1341.

Hedrick PW, 1985. Genetics of populations. Portola Valley, CA: Jones and Bartlett.

Hughes C, 1998. Integrating molecular techniques with field methods in studies of social behavior: a revolution results. Ecology 79:383–399.

Jones AG, Östlund-Nilsson S, and Avise JC, 1998. A microsatellite assessment of sneaked fertilizations and egg thievery in the fifteenspine stickleback. Evolution 52:848–858.

Kellogg KA, Markert J, Stauffer JR, and Kocher TD, 1998. Interspecific brood mixing and reduced polyandry in a maternal mouth-brooding cichlid. Behav Ecol 9:309–312.

Kichler K, Holder MT, Davis SK, Márquez MR, and Owens DW, 1999. Detection of multiple paternity in the Kemp's ridley sea turtle with limited sampling. Mol Ecol 8:819–830.

Levine L, Asmussen M, Olvera O, Powell JR, De La Rosa ME, Salceda VM, Gaso MI, Guzman J, and Anderson WW, 1980. Population genetics of Mexican *Drosophila.* V. A high rate of multiple insemination in a natural population of *Drosophila pseudoobscura.* Am Nat 116:493–503.

Marshall TC, Slate J, Kruulk LE, and Pemberton JM, 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol 7:639–655.

Parker A and Kornfield I, 1996. Polygynandry in *Pseudotropheus zebra,* a cichlid fish from Lake Malawi. Environ Biol Fish 47:345–352.

Taborsky M, 1994. Sneakers, satellites, and helpers—parasitic and cooperative behavior in fish reproduction. Adv Study Behav 23:1–100.

Corresponding Editor: John C. Avise