Title
Genetic parentage in large half-sib clutches: theoretical estimates and empirical appraisals.

Journal
Genetics, 154(4)

ISSN
0016-6731

Authors
DeWoody, J Andrew
Walker, DeEtte
Avise, John C

Publication Date
2000-04-01

DOI
10.1093/genetics/154.4.1907

Genetic Parentage in Large Half-Sib Clutches: Theoretical Estimates and Empirical Appraisals

J. Andrew DeWoody, DeEtte Walker and John C. Avise

Department of Genetics, University of Georgia, Athens, Georgia 30602

Manuscript received September 2, 1999
Accepted for publication December 23, 1999

ABSTRACT

Nearly all of the 906 embryos from a male-tended nest of the sand goby (*Pomatoschistus minutus*) were genotyped at two hypervariable microsatellite loci to document conclusively the number of mothers and their relative genetic contributions to the nest. The true number of mothers determined by this nearly exhaustive genetic appraisal was compared to computer simulation treatments based on allele frequencies in the population, assumptions about reproductive skew, and statistical sampling strategies of progeny subsets. The "ground-truthed" appraisal and the theoretical estimates showed good agreement, indicating that for this nest a random sample of ~20 offspring would have sufficed for assessing the true number of biological parents (but not necessarily their relative genetic contributions). Also, a general dilocus matrix procedure is suggested for organizing and interpreting otherwise cumbersome data sets when extremely large numbers of full-sib and half-sib embryos from a nest are genotyped at two or more hypervariable loci.

IN many fish (and other) species, molecular analyses of genetic parentage are made challenging by exceptionally large clutch sizes and numerous extrapair fertilizations. Unlike the situation in mammals or birds, for example, a fish clutch often consists of hundreds or thousands of eggs, and embryos in a single nest may stem from several or many biological parents (Taborsky 1994). To reconstruct the mating system, a complete genetic characterization of all embryos is neither practical nor necessary. Given that empirical appraisals of paternity and maternity in nest-guarding fish often come from a random sample of progeny, a question arises as to how large the sample sizes per nest must be to document parentage patterns adequately.

Computer simulations previously have been developed that estimate the sample sizes necessary for parentage assessment from a large clutch and the number of reproductive adults contributing to a nest (DeWoody *et al.* 2000a). These theoretical treatments make particular assumptions about reproductive skew (or lack thereof) in parental contributions to a nest. However, they have not been "ground-truthed" against empirical molecular data, a shortcoming that we begin to rectify here.

To assess parentage in a large nest directly, we have genotyped at two hypervariable microsatellite loci nearly all of 906 fish embryos from a nest of the sand goby, *Pomatoschistus minutus.* These genetic data (plus those from an additional 40 nests assayed less exhaustively) were used to (1) elucidate maternity and paternity, including the number of biological parents and their relative contributions to an offspring pool; (2) evaluate against empirical experience the theoretical predictions described in DeWoody *et al.* (2000a); (3) address sampling issues peculiar to large, mixed-maternity cohorts; and (4) introduce a means of visually organizing and displaying genetic data that can be informative about the parental gametotypes represented in large arrays of full-sib and half-sib progeny.

Corresponding author: J. Andrew DeWoody, Department of Genetics, Life Sciences Bldg., University of Georgia, Athens, GA 30602.
E-mail: dewoody@arches.uga.edu

MATERIALS AND METHODS

Microsatellite analyses: Sand goby collections and parentage analyses were conducted (A. G. Jones, D. W. Walker, C. Kvarnemo, K. Lindstrom and J. C. Avise, unpublished data). To briefly summarize those findings, embryos were genotyped at two microsatellite loci (*Sg1* and *Sgob5*), which displayed a total of 96 and 87 different alleles, respectively. From a subset of presumably unrelated adults assayed initially, these two loci yielded a combined parentage exclusion probability >0.995 (Equation 2a in Jamieson and Taylor 1997). Thus, the probability of excluding unrelated adults as potential parents of an embryo was very high. A third locus (*Sgob10*) was also typed in the few embryos whose parentage remained unresolved by the first two loci. Among 589 individuals assayed, 31 different alleles were detected at this locus (A. G. Jones, D. W. Walker, C. Kvarnemo, K. Lindstrom and J. C. Avise, unpublished data).

Of the 41 nests originally studied, one (TV12) was chosen here for exhaustive sampling efforts. We chose this nest because preliminary genetic analyses indicated that it was sired exclusively by one attendant male, and all progeny were heterozygous. Thus, the maternal alleles in each embryo were evident by subtraction. Each embryo thereby could be assigned a maternal "gametotype" (either an allele at one locus or a dilocus haplotype, depending on the context) that registers the gametic contribution of its mother. The minimum number

of mothers contributing to a nest was apparent either by counting maternal alleles at the most polymorphic locus (*e.g.*, Kellogg *et al.* 1995; DeWoody *et al.* 1998, 2000a) or by using the dilocus haplotype data. In principle, the maximum number of multilocus haplotypes produced by n mothers is equal to $2^L(n)$, where L is the number of loci assayed.

The number of embryos produced by different mothers was tallied and used to determine reproductive skew, in this case the relative maternal contributions to nest TV12. For each full-sib cohort within the nest, deviations of genotypic counts from Mendelian expectations were evaluated by chi square tests.

Rarefaction analyses: Both analytical and resampling procedures were used to generate rarefaction curves (Sanders 1968; Hulbert 1971; Heck *et al.* 1975; James and Rathbun 1981; Gotelli and Graves 1996). In the current application, these curves describe the number of different maternal gametotypes identified within a nest as an increasing function of the number of embryos sampled. The analytical method yields the expected number of maternal gametotypes in a random sample of n individuals drawn without replacement from a nest of N individuals, $E(H_n)$. This is given by

$$\sum_{i=1}^{H}\left[1 - \frac{\binom{N - N_i}{n}}{\binom{N}{n}}\right],$$

where H is the number of gametotypes in the nest and N_i is the number of individuals with haplotype i.

The resampling or randomization procedure for generating a rarefaction curve is analogous to bootstrapping. All genetically assayed embryos from a given nest were numbered sequentially and the first empirical appearance of each new gametotype in the list was recorded. A new random number was then assigned to each embryo, the numbers were sorted ascendingly, and again the first occurrence of each new gametotype was recorded. This process, repeated in our case 1000 times, yielded the mean number of sampled embryos required to detect all the distinct maternal gametotypes within a nest.

For example, in the empirical data set for nest TV12, in the 200th randomization trial we detected the sixth maternal allele at locus *Sgob5* in the 11th sampled embryo. In trial number 201, a sixth maternal allele was detected in the 25th embryo sampled. Note that this maternal allele is not necessarily identical in state between trials but rather is the sixth different allele encountered. An accumulation of such outcomes generated each rarefaction curve by the resampling method.

Computer simulations: *A priori*, we used the computer program BROOD (DeWoody *et al.* 2000a) to determine the mean sample size required to detect all maternal alleles within a nest. This estimator, if accurate, should coincide approximately with the plateau point in a rarefaction curve. To summarize the model, a simulated adult population was created using Hardy-Weinberg equilibrium and the empirically determined allele frequencies for the loci *Sg1* and *Sgob5*. Parents were then chosen at random from this population, and progeny arrays were constructed based upon the provisional assumption that seven or fewer mothers all contributed equally to the total pool of progeny within a nest. Progeny were sampled sequentially until all parental gametotypes were detected. This entire process was then repeated several thousand times in total.

Similarly, we also used the programs GAMETES and HAPLOTYPES to estimate the number of females contributing to a nest (DeWoody *et al.* 2000a,b). These resampling models create distributions of the number of gametotypes detected in a sample from a nest. Distributions were constructed for various parental "assemblage" sizes, where each assemblage consisted of the one shared parent and x unshared parents, where x assumes all integer values between one and some explicit maximum. The relative likelihood of each parental assemblage size was then determined and these estimates were compared to the empirical data.

RESULTS

Microsatellite analyses: Preliminary sampling suggested that sand goby nest TV12 was sired by a single male and, thus, consisted of a large collection of full-sib and half-sib progeny. Exhaustive genotyping supported this notion: all 864 embryos assayed at both loci (>95% of all embryos in nest TV12) possessed one of the expected paternal gametotypes. There was no evidence at these loci of either cuckoldry or *de novo* mutation (Jones *et al.* 1999).

By genotyping such a large cohort of progeny, the genotypes of all contributing mothers could be reconstructed accurately (Table 1). The highly nonrandom associations among maternal alleles at the two loci show that at least three different mothers were involved (Table 1). They produced ~47, 32, and 21%, respectively, of the assayed progeny in nest TV12.

The final entry in Table 1 (lower right corner) suggests that a fourth mother may have contributed six embryos to the nest. This evidence is puzzling, however, because only one (rather than two) maternal allele at each locus was detected in this small array of offspring. To assess the possibility of scoring errors or other anomalies, we also genotyped the embryos at a third locus (*Sgob10*). This did not fully resolve the puzzle, however,

TABLE 1

The distribution of maternal gametotypes among 864 embryos from sand goby nest TV12

	Sgob5						
Sg1	131	133	161	175	153	185	197
217	108	102					
233	114	81					
247			39	44			
251			45	52			
235					63	73	
275					57	80	
237							6

Alleles listed across the top are from locus *Sgob5*; those down the left side are from locus *Sg1*. In no case (with the exception of the lower right cell; see text) did a chi square test detect a significant departure of the dilocus gametotypic counts from Mendelian expectations within the progeny array of a given female.

When the data are arranged in this fashion, note the evident major contributions to the nest by three females, whose reconstructed dilocus gametic contributions at *Sgob5*:*Sg1* are as follows: 131/133:217/233; 161/175:247/251; and 153/185:235/275.

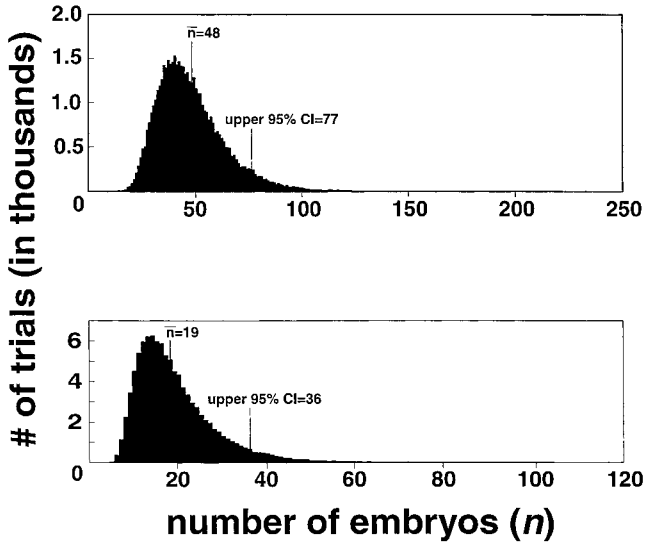

Figure 1.—Output from 1000 computer simulations (BROOD program in DeWoody *et al.* 2000a) estimating the sample sizes (n) of embryos needed to detect all maternal contributions to a nest, on the basis of the empirical levels of variation observed in sand goby populations at the microsatellite loci employed (see text). (Top) The *a priori* estimates (mean $n = 48$) assuming that no more than seven mothers contributed to a nest, all equally. (Bottom) The *a posteriori* estimates (mean $n = 19$) assuming the parameters empirically deduced for 864 embryos in nest TV12 (namely, three mothers with relative progeny contributions of 0.47, 0.32, and 0.21).

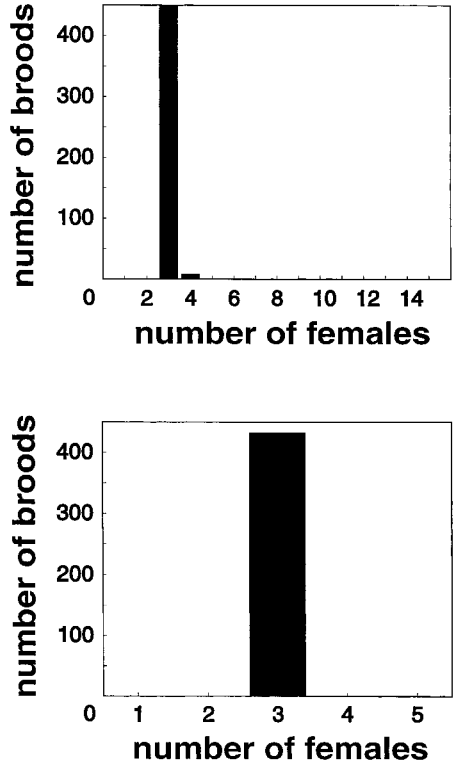

Figure 2.—Results of the GAMETES (top) and HAPLOTYPES (bottom) computer programs. As the number of unshared parents was varied systematically, the number of gametotypes was tabulated across many (several thousand in total) simulated nests. Shown are the number of mothers estimated by these programs for simulated nests whose embryos collectively display exactly six single-locus maternal gametotypes (top) and 12 dilocus maternal gametotypes (bottom), as did nest TV12. Of the >400 simulated nests that were composed of exactly 6 gametotypes (top), >95% were mothered by three dams. Furthermore, 100% of the >400 nests that were composed of exactly 12 gametotypes (bottom) were mothered by three dams. Thus, both programs indicate that the embryos in nest TV12 almost certainly came from three different mothers.

because in each case the maternal alleles were consistent with the genotypes of the other three known mothers (data not shown). In any event, if a fourth mother was involved, she contributed only six embryos to the nest, or 0.7% of the total.

Computer simulations and rarefaction: *A priori*, the computer simulations utilizing the BROOD program (DeWoody *et al.* 2000a) had suggested the need to sample an average of ~48 embryos (Figure 1, top; upper 95% confidence limit = 77) to detect gametotypes from as many as seven mothers assumed to have contributed equally to a typical sand goby nest, given the level of microsatellite polymorphism documented in the fish populations assayed. In practice, this calculation helped to guide our broader empirical study on sand gobies (A. G. Jones, D. W. Walker, C. Kvarnemo, K. Lindstrom and J. C. Avise, unpublished data).

For the three mothers documented with regard to nest TV12, we repeated the BROOD simulations *a posteriori*, first using a uniform distribution of reproductive success and then using the observed values (*i.e.*, 0.47, 0.32, and 0.21). This time, the simulations decreased the mean sample size necessary to detect all contributing mothers for that particular nest to 17 and 19 embryos, respectively (observed values shown in Figure 1, bottom; upper 95% confidence limit = 36). The GAMETES and HAPLOTYPES simulations (Figure 2) each clearly indicated that three mothers contributed to nest TV12. Indeed, the likelihood of three mothers was >40 times greater than that of four mothers, the next most probable outcome.

Based again on the empirical genetic data for TV12, rarefaction curves summarizing the mean sampling effort required to detect maternal gametotypes are compared in Figure 3. Note the close agreement between the rarefaction curves generated by the analytical and by the resampling methods.

DISCUSSION

Empirical data: A preliminary subsample of embryos from nest TV12 suggested that all of the progeny were sired by the nest-attendant male. However, the possibility remained that more extensive sampling might reveal a low level of fertilization thievery, especially given the high frequency of cuckoldry by sand goby males (A. G. Jones, D. W. Walker, C. Kvarnemo, K. Lindstrom and

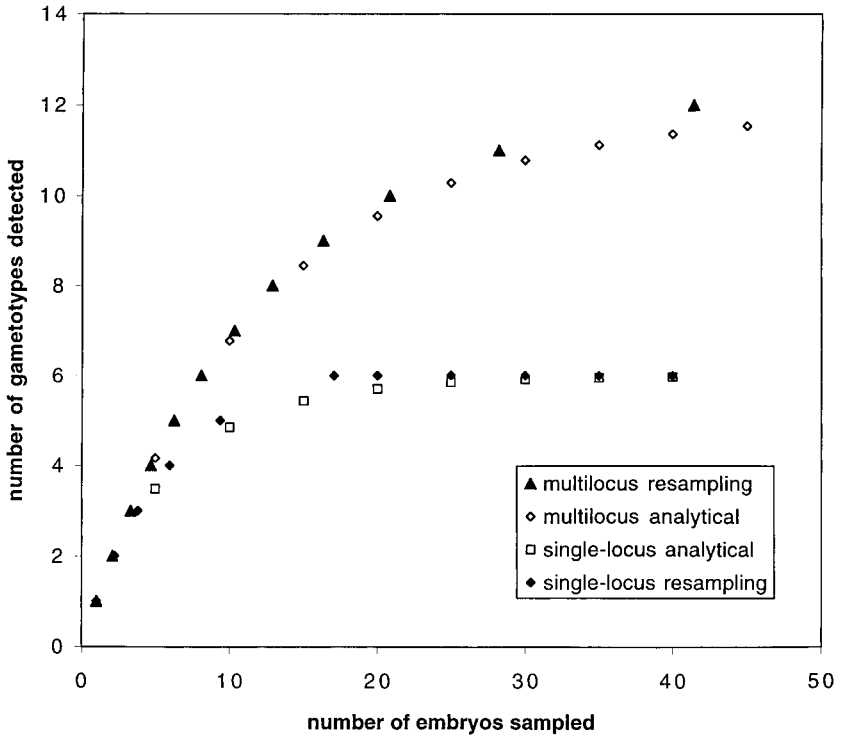

Figure 3.—Rarefaction curves showing the estimated mean number of embryos needed to detect maternal gametotypes in nest TV12. The simulated curves were generated by the resampling procedure (1000 bootstrap replicates); analytical curves were calculated explicitly.

J. C. Avise, unpublished data). The subsequent, near-exhaustive sampling of embryos from nest TV12 proved that the attendant male in this case was not cuckolded. This in turn facilitated the genetic maternity analysis. Most sand goby nests are known to contain eggs and embryos from several females (Forsgren *et al.* 1996; (A. G. Jones, D. W. Walker, C. Kvarnemo, K. Lindstrom and J. C. Avise, unpublished data). Nest TV12, with at least three firmly documented mothers for its 864 embryos, proved to be unexceptional in this regard.

We have no satisfying explanation for the six genetically anomalous embryos in nest TV12 (Table 1, bottom right). *De novo* mutations, either arising alone (DeWoody *et al.* 1998) or as premeiotic clusters (Woodruff and Thompson 1992; Jones *et al.* 1999), seem an improbable source given that a novel allele was present at both loci simultaneously. Perhaps this genotype registers the contribution of a fourth mother to the nest, but this explanation, too, seems somewhat strained because it would require either that she was an (extremely rare) double homozygote or that she was a normal double heterozygote who happened to transmit only one of four (presumably equally likely) dilocus genotypes to all six of her offspring. Other potential explanations invoke nonamplifying or "null" alleles (not otherwise seen in this study) at both loci, shared maternal alleles with other known mothers of this nest, species misidentifications of these six embryos, or some combination of the above possibilities. In any event, $<1\%$ of the embryos in nest TV12 were problematic, so for the sake of simplicity we assume in the following discussion that nest TV12 had only the three firmly documented mothers.

Empirical appraisal of the statistical methods: The *a priori* BROOD simulations suggested that on average a random sample of $\sim n = 48$ embryos would be necessary and sufficient to detect all maternal alleles in a half-sib progeny array of sand gobies (Figure 1, top), provided that no more than seven mothers contributed to a nest, all equally. By contrast, rarefaction analyses as applied to the empirical data for TV12 indicated that the sample size for that nest could have been considerably smaller.

The *a posteriori* BROOD simulations for nest TV12 (Figure 1, bottom) also indicate that the *a priori* BROOD estimates in that case were inflated. In the *a priori* simulations, we (conservatively) used a presumed maximum possible number of mothers who might have contributed to a nest, rather than the true number, which only became apparent later. When higher numbers of mothers are permitted in the simulations, more embryos per nest must be sampled (all else being equal) to ensure detection of all maternal gametes. After correcting the BROOD simulations for the true maternal number, as determined from exhaustive genetic analysis, the *a posteriori* estimate of the number of embryos ($n \cong 19$ using the observed reproductive skew and $n \cong 17$ assuming uniform maternal contributions) that would suffice as an adequate sample from nest TV12 matched closely the estimate from empirical rarefaction ($n \cong 17$).

Note that this discussion pertains to the mean number of embryos to be sampled from an average nest. In practice, depending on the goals of the study, the costs and benefits of surveying additional embryos should be considered. For example, if the intent is to identify every

A

	100	102	104	106	108	110
200	75	75				
202	75	75				
204			75	75		
206			75	75		
208					75	75
210					75	75

B

	100	102	104	106	108	110
200	75	75				
202	75	75				
204			75	75		
206			75	75		
208				75	75	
210				75	75	

C

	100	102	104	106	108	110	112
200	75	75					
202	75	75					
204			75	75			
206			75	75			
208					75	74	1
210					75	75	

D

	100	102	104	106	108	110	112
200	75	75					
202	75	75					
204			75	75			
206			75	75			
208					75	70	5
210					75	70	5

E

	100	102	104	106	108	110
200	150					
202		150				
204			150			
206				150		
208					150	
210						150

F

	100	102	104	106	108	110
200	75	75				
202	75	75				
204			75	75		
206			75	75		
208					100	50
210					100	50

Figure 4.—Dilocus tables illustrating how different genetic phenomena can leave distinct "signatures" on allelic associations in a nest of 900 embryos. (A) A null model where three different mothers (all double heterozygotes with no alleles shared) contributed equally to the nest; (B) an allele (106) was shared by two mothers; (C) a singleton *de novo* mutation (to allele 112) occurred at one locus; (D) a premeiotic clustered mutation (to allele 112) occurred at one locus; (E) complete linkage disequilibria; and (F) meiotic drive by allele 108.

gametotype in each nest (which might be the case in a quantitative study of meiotic drive, for example), then the more appropriate sample sizes of embryos per nest might be estimated as the upper 95% confidence interval (Figure 1), as opposed to the mean.

As mentioned previously, the BROOD simulations provisionally assume that all mothers contribute equally to the pool of embryos within a nest. For nest TV12, the observed relative contributions of the three mothers (0.47, 0.32, and 0.21) were not grossly unequal. Apparently for this reason, nearly all of the difference in the *a priori vs. a posteriori* sample sizes in BROOD was due to the decrease in the number of mothers in the simulations (from seven to three), rather than to the modest observed departure from a uniform distribution of female contributions to this nest. To our knowledge, there are no other reliable genetic estimates of reproductive skew in extremely large nests of externally fertilizing fishes, so clearly this is an important area for further exploration on both the empirical and theoretical fronts. If reproductive skew typically is more pronounced in most nests, empirical sample sizes needed to detect all mothers will have to be larger to ensure capture of the rarer maternal contributions. In agreement with the exhaustive empirical data (Table 1), the GAMETES and HAPLOTYPE programs both estimated that three mothers contributed to nest TV12 (Figure 2). Thus, the observed modest departure from equal maternal contributions to this nest did not bias the estimated number of mothers appreciably.

Both empirical rarefaction analyses and theoretical simulations can be used to develop appropriate strategies for genetic sampling (Kohn *et al.* 1999). As typically applied, the rarefaction method is specific for a population from which samples were taken. Thus, in the current context, asymptotes in the rarefaction curves are nest specific and may differ from case to case depending upon the number of females contributing to a progeny array, the genotypes of those mothers, and the reproductive skew.

The BROOD approach, on the other hand, estimates mean sample sizes on the basis of many different nests. Huge numbers of nests are generated and each is sampled hundreds or thousands of times to produce confidence intervals around mean sample sizes needed to detect all parents given a specified set of polymorphic markers. The current rarefaction analyses of empirical data for the sand goby suggest, at least in this case, that the simulations perform reasonably well.

The computer simulations presented in DeWoody *et al.* (2000a) require half-sib progeny arrays, as was the case with nest TV12. However, many species generate

progeny arrays that are composed of full-sibs ($r = 0.5$), half-sibs ($r = 0.25$), or even unrelated individuals ($r = 0.0$). For example, if group spawning (involving multiple males and females) occurs, embryos of varying degrees of relatedness will be produced. Clearly, more work is needed to evaluate sampling concerns in these more complicated situations.

Matrix representations of progeny data from large nests: When large numbers of embryos from multiple mothers are sampled from a nest at two or more hypervariable loci, data management can be cumbersome. We have found that by organizing the information into dilocus matrices of maternal allelic associations (as in Table 1), informative patterns emerge that otherwise may be less apparent.

Figure 4 illustrates the hypothetical signatures of six different genetic phenomena that can become evident when the deduced maternal gametotypes of full-sib and half-sib embryos are tabulated in this manner. The null model (Figure 4A) is that several mothers shared no alleles and contributed equally to a progeny array. Various patterns of departure from the null are shown in the other panels. For example, Figure 4B shows the expected departure from the null pattern due to the effects of two mothers having shared a marker allele at one locus. Figure 4, C and D, shows "orphan" genotypic entries into such a matrix that would signify the presence of late-meiotic singleton mutations and premeiotic mutational clusters (Jones *et al.* 1999), respectively, within the nest. Figure 4, E and F, illustrates the effects of gametic-phase disequilibrium and meiotic drive, respectively.

In principle, such representations could be extended to three (or more) loci simultaneously by plotting the gametotypic data in more than two dimensions. However, vastly larger numbers of cells would be entailed in such treatments, and, hence, genotypic counts in most cells inevitably would be too low for statistical assessment even in the largest of nests normally available for empirical studies. For these reasons, and also because two hypervariable loci normally afford more than adequate exclusionary power in assessments of genetic parentage, the dilocus representations should normally suffice.

We thank A. Fiumera, A. Jones, C. Kvarnemo, A. Keyser, K. Lindstrom, M. Mackieowitz, B. McCoy, D. Pearse, B. Porter, and D. Promislow for their input and support. Work was funded by the Pew Foundation and the University of Georgia.

LITERATURE CITED

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

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. Fletcher, S. D. Wilkins and J. C. Avise, 2000b Parentage and nest guarding in the tessellated darter (*Etheostoma olmstedi;* Perciformes) assayed by microsatellite markers. Copeia (in press).

Forsgren, E., C. Kvarnemo and K. Lindström, 1996 Mode of sexual selection determined by resource abundance in two sand goby populations. Evolution **50:** 646–654.

Gotelli, N. J., and G. R. Graves, 1996 *Null Models in Ecology*. Smithsonian Press, Washington, DC.

Heck, K. L., G. van Belle and D. Simberloff, 1975 Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology **56:** 1459–1461.

Hulbert, S. H., 1971 The non-concept of species diversity: a critique and alternative parameters. Ecology **52:** 577–586.

James, F. C., and S. Rathbun, 1981 Rarefaction, relative abundance, and diversity of avian communities. Auk **98:** 785–800.

Jamieson, A., and S. Taylor, 1997 Comparisons of three probability formulae for parentage exclusion. Anim. Genet. **28:** 397–400.

Jones, A. G., G. Rosenqvist, A. Berglund and J. C. Avise, 1999 Clustered microsatellite mutations in the pipefish *Syngnathus typhle*. Genetics **152:** 1057–1063.

Kellogg, K. A., J. A. Markert, J. R. Stauffer, Jr., and T. D. Kocher, 1995 Microsatellite variation demonstrates multiple paternity in lekking cichlid fishes from Lake Malawi, Africa. Proc. R. Soc. Lond. Ser. B **260:** 79–84.

Kohn, M. H., E. C. York, D. A. Kamradt, G. Haught, R. M. Sauvajoy *et al.*, 1999 Estimating population size by genotyping faeces. Proc. R. Soc. Lond. Ser. B **266:** 657–663.

Sanders, H. L., 1968 Marine benthic diversity: a comparative study. Am. Nat. **102:** 243–282.

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

Woodruff, R. C., and J. N. Thompson, 1992 Have premeiotic clusters of mutation been overlooked in evolutionary theory? J. Evol. Biol. **5:** 457–464.

Communicating editor: Z-B. Zeng