Bayesian Sperm Competition Estimates

Beatrix Jones*,[1] and Andrew G. Clark†

**Department of Statistics, Penn State University, University Park, Pennsylvania 16802 and †Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853*

Manuscript received July 24, 2002
Accepted for publication December 12, 2002

ABSTRACT

We introduce a Bayesian method for estimating parameters for a model of multiple mating and sperm displacement from genotype counts of brood-structured data. The model is initially targeted for *Drosophila melanogaster*, but is easily adapted to other organisms. The method is appropriate for use with field studies where the number of mates and the genotypes of the mates cannot be controlled, but where unlinked markers have been collected for a set of females and a sample of their offspring. Advantages over previous approaches include full use of multilocus information and the ability to cope appropriately with missing data and ambiguities about which alleles are maternally *vs.* paternally inherited. The advantages of including X-linked markers are also demonstrated.

SPERM competition is an important factor in the evolution of reproduction of many organisms, particularly birds and insects. The phenomenon of sperm competition in Drosophila has been extensively studied by setting up matings in the laboratory with males that bear different visible genetic markers (Fowler 1973; Prout and Bundgaard 1977). Females mate with multiple males and store the collection of sperm in the seminal receptacle and the paired spermathecae for later use in fertilizing eggs. These laboratory experiments have shown that later-mating males tend to father a greater proportion of the offspring and that there is great variability among genotypes of males and of females in the magnitude of this later-male advantage (Clark *et al.* 1995; Clark and Begun 1998). The precise mechanism that determines sperm success is not fully understood; however, by following fluorescently labeled sperm Civetta (1999) and Price *et al.* (1999) showed that the sperm that gets transported to the storage organs will be used in fertilization. This implies that the critical time is shortly after mating, when the decision is made as to which sperm will be stored. When the female remates, it appears that there is some loss of the first male's sperm, through either physical removal or incapacitation (Price *et al.* 1999), so the phenomenon is often called sperm displacement.

There has always been some concern that laboratory experiments in sperm competition are somewhat contrived in that the females must be exposed to males in a prescribed order and timing. The timing of opportunities for mating in natural populations no doubt differs dramatically from such laboratory experiments, so the whole phenomenon of sperm competition may be quite different in nature. It would be desirable to study sperm competition directly by sampling from natural populations. However, this presents more challenges than the laboratory setting. In natural populations, typically the mother and a sample of the offspring (the brood) are genotyped at one or more loci, but no information is available on the fathers, except perhaps population allele frequencies. However, it is still possible to use the available information to model the number of mates for each mother and to quantify attributes of sperm competition.

Harshman and Clark (1998) developed a model for multiple mating and sperm displacement, building on the approach of Cobbs (1977) and Griffiths *et al.* (1982). To fit the parameters of the model for data from a wild population, they reduced the single-locus data for each brood to a summary statistic, the number of distinct paternal alleles. They then used simulation to estimate the distribution of this statistic for a grid of parameter values. The likelihood of a set of parameter values for a particular brood was then simply the height of the simulated distribution at the observed number of distinct paternal alleles at each locus. Information from multiple loci was combined by taking the product of likelihoods across loci.

Because this approach does not use the allele counts, differentiate between alleles of different frequencies, or use haplotype information in the multilocus case (*i.e.*, is not based on sufficient statistics) some information is lost. Computational constraints also led Harshman and Clark to consider at most four possible mates per brood. We take the model from Harshman and Clark (1998), place it in a Bayesian framework, and use Markov chain Monte Carlo to examine the posterior distribution of the model parameters. This approach allows us to make full use of the data, remove limits on the

[1]*Corresponding author:* 326 Thomas Bldg., Department of Statistics, Penn State University, University Park, PA 16802.
E-mail: trix@stat.psu.edu

number of possible mates per brood, and deal with missing data and ambiguities in which alleles are paternally inherited. We examine the performance of this method on both simulated data and the data considered in Harshman and Clark (1998).

METHODS

Model for mating and offspring production: The model of sperm competition is that each mate following the first mate displaces fraction β of the already-present sperm. If there are K mates, the first mating male accounts for fraction $(1 - \beta)^{K-1}$ of the stored sperm; the ith mating male accounts for $\beta(1 - \beta)^{K-i}$. Laboratory experiments with just two mates have consistently shown that the later-mating male fathers more offspring, so $\beta > 0.5$ is a biologically plausible constraint to place on the estimator. Thus we have put a prior on β that is uniform between 0.5 and 1.

The number of sampled offspring fathered by each mate is assumed to follow a multinomial distribution, where the multinomial probabilities correspond to the fraction of sperm from each father. The counts of offspring with different paternal haplotypes are thus from a different multinomial, where the probability of each genotype is a sum over fathers, weighted by the fraction of sperm from that father. For a haplotype h that segregates from the ith mate with probability x_{ih}, the probability is

$$p_h = (1 - \beta)^{K-1}x_{1h} + \sum_{i=2}^{K}\beta(1 - \beta)^{K-i}x_{ih}. \qquad (1)$$

Of course, the x's can be computed only when the number of mates for each brood (K's), the genotypes of each mate (g_i), and the paternal alleles of the offspring (a's) are known. We treat these as nuisance parameters; our approach is to sample from the joint posterior of α, β, the K's, a's, and g's and then marginalize to get the posterior of α and β.

The number of mates K is assumed to have a truncated Poisson distribution (since all the females produced offspring, the possibility of zero mates has been eliminated). This distribution is parameterized by α:

$$\pi_K(k) = \Pr(K = k) = \frac{\alpha^k e^{-\alpha}/k!}{1 - e^{-\alpha}}. \qquad (2)$$

This equation can be thought of as a prior on the value of K for each brood. We place a uniform (1, 6) prior on α. An upper bound of 6 in the prior is arbitrary, although in practice it would be difficult to detect this many distinct inferred fathers among the offspring.

The prior on the mates' genotypes is just the population frequencies (assuming Hardy-Weinberg and linkage equilibrium). Call this $\pi_G(g)$. The population allele frequencies are assumed to be known, but in reality are estimated from the data. The estimates used are the observed allele frequencies among the mothers' alleles and the offspring's paternal alleles. For purposes of allele frequency estimation only, if there is ambiguity in which the allele is paternal, the paternal allele is arbitrarily designated. It would also be possible to sample adult males from the population and to compare the allele frequencies to those inferred from the progeny to infer differential mating success (Bundgaard and Christiansen 1972), but for now we assume that no adult male data are available.

In many cases, the probability that offspring i receives allele a_{il} at locus l through paternal inheritance, $\pi_A(a_{il})$, is simply 0 or 1: The paternal allele received by an offspring is ambiguous only if the offspring shares both of its alleles with its mother at that locus. If there are ambiguities, π_A places half its mass on each of the two possible alleles. Thus the probability of any feasible set of values for a is just $(\frac{1}{2})^{\text{Number of ambiguities}}$.

The posterior probability of a particular set of (α, β, K, g, a) is then proportional to the prior times the likelihood:

$$P(\alpha, \beta, k, g, a) \propto \prod_{i=0}^{\text{No. broods}} \text{Mult}(N_i, p_i)\pi_K(K_i)\pi_G(g_i)\pi_A(a_i).$$

The normalizing constant for this distribution is unknown, so Markov chain Monte Carlo is used to characterize its properties.

Markov chain Monte Carlo parameter estimation: We construct an ergodic Markov chain whose stationary distribution is the posterior of α, β, the K's, a's, and g's. A sample from this chain can thus be used to estimate the general shape of the posterior distribution, its mode, mean, and other quantities of interest.

The chain moves among the possible values for α, β, the K's, a's, and g's. It is constructed using a reversible jump Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970; Green 1995). Reversible jump simply refers to the fact that when the number of fathers changes, the dimension of g does as well. However, the Jacobian of the dimension "jumping" transformation is one, so the algorithm is identical to a standard Metropolis-Hastings. Moves are proposed, and then the associated Hastings ratio is computed:

$$\frac{P([\alpha, \beta, k, g, a]')q([\alpha, \beta, K, g, a], [\alpha, \beta, K, g, a]')}{P([\alpha, \beta, K, g, a])q([\alpha, \beta, K, g, a]', [\alpha, \beta, K, g, a])}.$$

This is the ratio of the posterior probabilities, adjusted for asymmetries in the proposal distribution: $q([\alpha, \beta, K, g, a]', [\alpha, \beta, K, g, a])$ is the probability of proposing to move to $[\alpha, \beta, K, g, a]'$ when the current state is $[\alpha, \beta, K, g, a]$; $q([\alpha, \beta, K, g, a], [\alpha, \beta, K, g, a]')$ is the probability of proposing the reverse move. A uniform (0, 1) random number is generated; if it is less than the Hastings ratio, the move is accepted (note that this means proposals with a Hastings ratio larger than one are always accepted). If a move is rejected, a second sample at the current state is recorded.

The moves we use to sample the possible state for each brood were:

1. Change the *i*th father's genotype at some locus *l*. Both *i* and *l* are chosen uniformly from the possible values.
2. Change the order of the fathers. Two fathers were selected at random for swapping.
3. Add a father. A new first-mating father is proposed; the alleles for this father are selected with equal probability for each paternal allele appearing in the brood and 1 allele representing all others in the population. [Each allele is proposed with probability 1/(number of observed alleles in this brood + 1).]
4. Subtract a father. This move proposes to delete the first-mating father.
5. Switch ambiguous alleles. We proposed to switch the allele designated as paternally inherited, a_{ib}, from one of the offspring's alleles to the other. Simultaneously, we propose to switch values of the fathers' genotypes matching a_{il}'s original value. We propose to switch each instance with probability 0.5.

The broods are cycled through; for each brood one of the preceding moves is proposed, each with probability 0.2. After a cycle through the broods, α and β are updated. A new α is proposed by sampling uniformly from a window of width 1 centered on the current α; a similar mechanism is used to propose the new β, but the window has width 0.1.

The resulting chain is irreducible (each state is reachable from every other state, and the number of steps between two states need not be a multiple of any number greater than one) and thus samples from the desired posterior. This algorithm has been coded in C++ as SCARE (sperm competition and remating estimates), available from http://www.stat.psu.edu/~trix/software/scare.html.

Simulation study: The performance of the method was examined by simulating 100 data sets under each of nine different scenarios and by assessing the accuracy of the inferences made. The first three scenarios each involve typing a total of 400 offspring, the next three a total of 900, and the final three a total of 1600. The first three scenarios each consisted of 20 broods with 20 typed offspring. Scenario 1 used one autosomal locus, scenario 2 used two autosomal loci, and scenario 3 used a single X-linked locus. (In this final case, it was presumed that only female offspring were sampled.) There were 15 equifrequent alleles per locus, the sort of polymorphism one might expect at microsatellite loci. (This information was not used in the inference procedure; the allele frequencies were estimated from each data set.) The remaining scenarios all used a single autosomal locus. A "square" design with 30 broods and 30 offspring per brood was compared with designs with 20 broods and 45 offspring per brood and 45 broods and 20 offspring per brood. Similarly, the scenarios with 1600 total offspring examined three designs with 40 broods of 40 offspring, 20 broods of 80 offspring, and 80 broods of 20 offspring.

For each scenario, broods were simulated by first sampling from the distribution of numbers of mates per female, with parameter $\alpha = 3$. The relative contribution of the male genotypes to the sperm pool was as specified by our model with $\beta = 0.7$. Over the range of plausible values of α (1–6) and β (0.6–0.9), the simulations perform less well for a given sample size as the number of mates increases and as the magnitude of sperm competition increases, although a systematic quantification of these effects was not done. For each simulated data set a sample of 10,000 parameters from the posterior distribution of (α, β) was generated using the SCARE software. The samples are spaced by 100 cycles through the broods. This number of samples was determined to be adequate because repeated runs of this length, even with different starting points, gave very similar parameter estimates for the Ravenswood data, as discussed below.

Ravenswood data: We consider the same data studied in Harshman and Clark (1998). The data were collected from a natural population of *Drosophila melanogaster* living near the open fermenters of the Ravenswood winery in Sonoma County, California. Female flies were captured and stored in vials. The vials were maintained at room temperature until the females laid eggs and the eggs hatched. Each female and a sample of her offspring were then genotyped for two microsatellite loci, *ula* and *nanos*. A total of 19 broods were collected, with an average of 13 offspring typed per brood. The two loci are actually linked, with a recombination fraction of ~20%, a deviation from the assumptions the SCARE software used in computing the haplotype probabilities (x_{ih}'s) in Equation 1. While it is conceptually straightforward to accommodate the linkage, we perform the estimation as if the loci are unlinked and discuss what effect this might have.

Again, inference based on a sample of 10,000 was taken from the joint posterior of α and β, spaced by 100 Metropolis-Hastings updates for each brood. The starting values for α and β were 2.0 and 0.6, respectively. To assess Monte Carlo error, this procedure was repeated 100 times with different random number seeds. The procedure was also tried with several different starting values for (α, β), spread over the region with prior support: (1.0, 0.5), (1.0, 0.9), (5.0, 0.5), and (5.0, 0.9).

RESULTS

Simulated data: Histograms of the posterior means for α inferred for the simulated data sets under each design scenario are in Figure 1. The posterior means for β are in Figure 2. The method shows a tendency to underestimate α when there are low numbers of offspring per brood and a single autosomal locus. Either

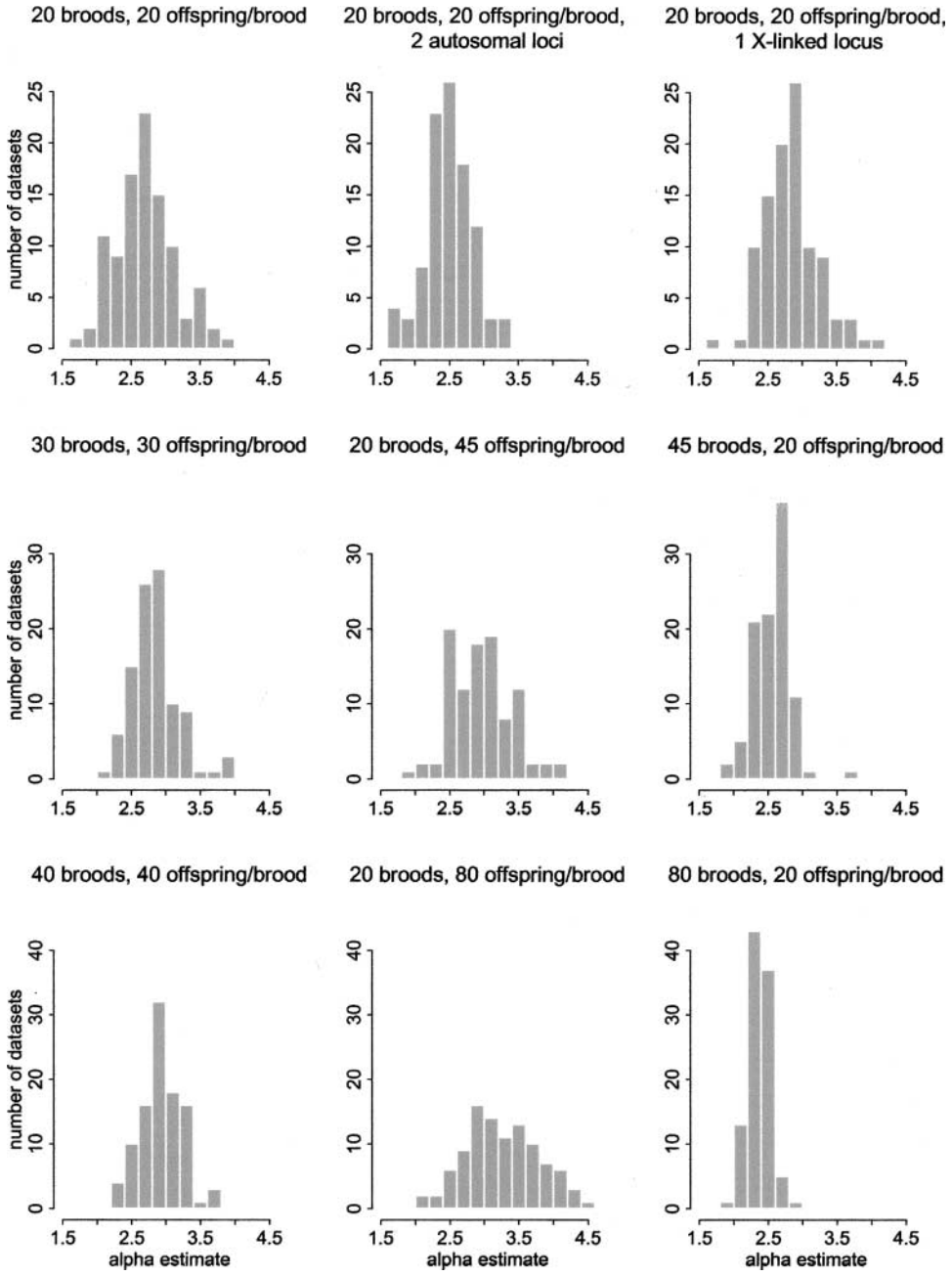

FIGURE 1.—Histograms of the posterior mean for α under different simulated design scenarios. Each histogram represents 100 data sets. One autosomal locus was used unless otherwise noted.

adding additional offspring or using an X-linked locus improves the situation.

β also tends to be slightly underestimated; although the range of the β estimates narrows as the total number of typed offspring increases, this bias seems to persist over the range of designs considered. Using an X-linked locus or two loci also improves estimation for β.

Ravenswood data: A histogram of 10,000 samples from the joint posterior of α and β is shown in Figure 3. From this sample, posterior means and credible intervals were estimated. For α, the posterior mean was 2.44 with a 90% credible interval of (1.64, 3.32); for β the posterior mean was 0.61 with a 90% credible interval of (0.51, 0.69). The Monte Carlo standard deviation for the posterior mean estimate, estimated by 100 runs with different random number seeds, was 0.021 for α and 0.0045 for β. As desired, the uncertainty due to Monte Carlo error is dwarfed by the parameter uncertainty (as represented by the credible intervals).

The estimates of posterior means were also insensitive to the starting values for the parameters. Three of the four "extreme" starting values produced estimates falling within the interquartile span of the 100 estimates using starting value (2.0, 0.6). The exception was starting value (5.0, 0.9), which, despite starting with a value at the high end for both parameters, wound up with a (somewhat) unusually low estimate for both: (2.39, 0.59). This estimate still falls within the range of the

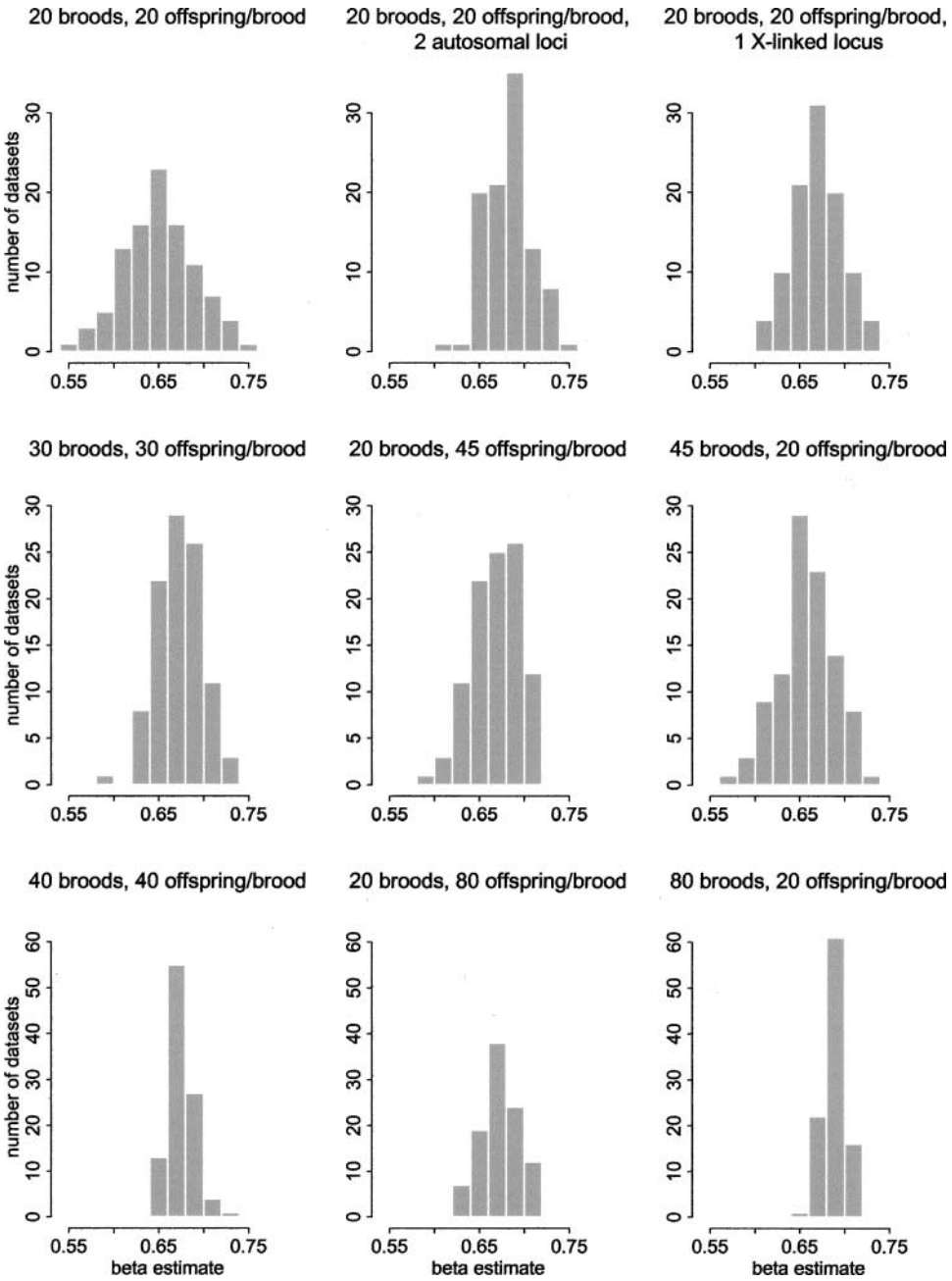

FIGURE 2.—Histograms of the posterior mean for β under different simulated design scenarios. Each histogram represents 100 data sets. One autosomal locus was used unless otherwise noted.

samples produced with starting point (2.0, 0.6) and clearly shows that there is no problem with "stickiness" at the starting value.

DISCUSSION

Simulated data: The tendency to underestimate α with low numbers of offspring reflects the fact that for these designs it is more likely for a mate to have no offspring among those typed. The likelihood tends to favor parsimonious configurations of the fathers' genotypes, so if a mate has no offspring among those sampled the estimate of K for that brood, and therefore the estimate of α, shifts downward. The problem of mates without offspring in the sample remains even when there is information that helps resolve paternity, such as using an X-linked rather than an autosomal locus or multiple loci. In fact, additional simulations (not shown) demonstrate that the pressure for parsimonious paternal configurations increases with the polymorphism of the locus, or the number of loci, as the probability of chance matches between additional fathers and the observed offspring decreases. However, as the number of offspring increases the bias subsides; thus we see that as more offspring per brood are included, the histograms become more balanced around the true α value of 3.0.

While increasing the number of offspring per brood

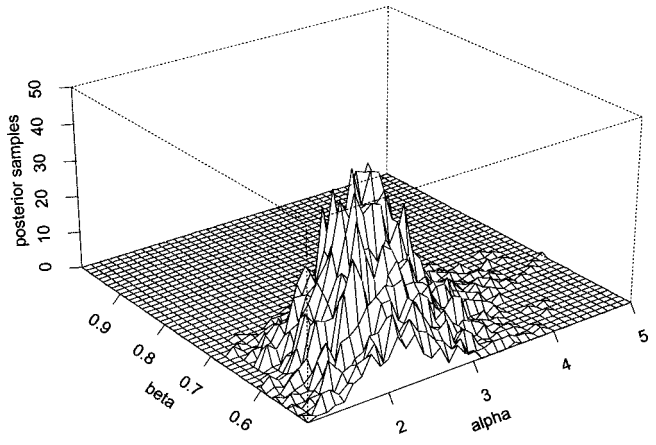

FIGURE 3.—Two-dimensional histogram of 10,000 samples from the posterior of α and β for the Ravenswood data.

helps reduce bias in α, since the distribution parameterized by α describes variation in the number of mates between broods, increasing the number of broods is necessary to decrease the variance of the estimate of α. This is reflected in the fact that although the "rectangular" designs with more offspring per brood are more symmetric around the true α value, they have a larger spread than designs that have the same total number of offspring but more broods. Consequently, it seems that a "square" design provides the best solution, balancing the effects of variance and bias.

In contrast, if assignment to paternal groups could be done without error, increasing either the number of offspring or the number of broods (while holding the other steady) would lead to improved estimation of β. This seems to continue to hold in our situation where we have imperfect parentage information: There are not marked differences in the distributions of β estimates when the same total numbers of typed offspring are used.

X-linked or multiple loci improve estimation of β by better resolving paternity. Our simulations show X-linked loci are effective at resolving alternative fathers, and thus improving estimation of β, without additionally biasing the estimate of α. Thus it seems that when only a small number of offspring can be typed, switching to use of an X-linked locus is a better way to improve the parentage information than using multiple autosomal loci. Once the number of offspring per brood is large enough to ensure a high probability of including offspring from all mates, we would ideally include enough genetic information (from a mix of X-linked and autosomal markers) that each offspring has a high probability of uniquely specifying a paternal genotype.

Laboratory research suggests that β in fact varies among males (CLARK *et al.* 1995). One goal might be to characterize this variation in a natural population. However, to do this the variation in the proportions of offspring with each genotype caused by differing β's must be large compared to the variation due to multinomial sampling of the offspring from each paternal group and ambiguity in assigning the paternal groups. The histogram with 40 broods and 40 offspring (1600 offspring total) reflects these sources of variation and has a range of ~0.07. Only if the variation of β between males exceeds this could we hope to detect it with such a design.

Ravenswood data: In this data set, we do not expect the fact that the loci are treated as independent to produce misleading results. The main effect of a deficiency of recombinant haplotypes is to reduce the information about which haplotypes are paired within a father: This information will be intermediate between that found in a data set with two independent loci and a data set with a single (more polymorphic) locus.

Allele frequencies were estimated for the data set as described in METHODS; it is worth noting that if one of the loci used was in linkage disequilibrium with a male locus affecting sperm competition, the allele frequencies for that locus could be substantially biased. However, we have no reason to believe that is the case for this data set. Although the frequencies are based on ~271 individuals (the 19 mothers and their offspring), the paternal alleles of offspring in the same brood are correlated so the variance of the estimate is much larger than that for 271 independent individuals. Estimation of the allele frequencies could be incorporated into the Markov chain Monte Carlo and utilizes the inferred paternal genotypes (the *g*'s) directly; an advantage of this approach would be that posterior distribution would reflect the uncertainty in α and β due to imperfect knowlege of the allele frequencies.

Our analysis of the Ravenswood data shows a lower value for β (0.61 *vs.* 0.83) and a higher value for α (2.44 *vs.* 1.82) than that reported in HARSHMAN and CLARK (1998). The value of α reported by Harshman and Clark is within our 90% credible interval, so in this sense it is not dramatically different; however, their estimate of β is above the credible interval obtained here. We believe the main reason for this is that the method described in this article makes full use of multilocus information to resolve differing paternity, while the method described in Harshman and Clark does not. Their likelihoods are based on the number of distinct alleles observed at each locus. Many broods in this data set, looked at in that manner, are compatible with a smaller number of fathers than is possible if the multilocus paternal haplotypes are considered. That clearly explains the increase in α; the decrease in β is explained by the fact that with the small brood sizes (13 on average) we would not expect offspring from so many fathers to end up in our sample unless β is relatively low (resulting in a more even distribution of offspring among mates). This substantial change highlights the importance of our methodology's ability to fully utilize multilocus information. Biologically, the finding that the degree of last-

male advantage gets him only 61% of the progeny may be quite important, because the assumption had been that the level of sperm competition in nature ought to be comparable to that in the laboratory, where more typically 90% or more of the progeny are sired by the last male. Of course, in the laboratory, typically only two successive males are tested, so the recurrent mating by females may serve to reduce the overall magnitude of sperm competition. This is contrary to the general belief that the higher the frequency of repeated matings, the stronger the opportunity for sperm competition should be. Clearly there is room to apply these methods to better understand the nature of sperm competition in field situations.

We thank Drs. Anthony Fiumera and Lawrence Harshman for helpful comments on a draft of this manuscript. This work is supported by National Science Foundation grant DEB 0108965 to A.G.C.

LITERATURE CITED

Bundgaard, J., and F. Christiansen, 1972 Dynamics of polymorphisms. I. Selection components in an experimental population of *Drosophila melanogaster*. Genetics **71:** 439–460.

Civetta, A., 1999 Direct visualization of sperm competition and sperm storage in Drosophila. Curr. Biol. **9:** 841–844.

Clark, A. G., and D. J. Begun, 1998 Female genotypes affect sperm displacement in Drosophila. Genetics **149:** 1487–1493.

Clark, A. G., M. Aguade, T. Prout, L. G. Harshman and C. H. Langley, 1995 Variation in sperm displacement and its association with accessory-gland protein loci in *Drosophila melanogaster*. Genetics **139:** 189–201.

Cobbs, G., 1977 Multiple insemination and male sexual selection in natural populations of *Drosophila pseudoobscura*. Am. Nat. **111:** 641–656.

Fowler, G. L., 1973 Some aspects of the reproductive biology of Drosophila: sperm transfer, sperm storage, and sperm utilization. Adv. Genet. **17:** 293–360.

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

Griffiths, R. C., S. McKechnie and J. A. McKenzie, 1982 Multiple mating and sperm displacement in natural populations of *Drosophila melanogaster*. Theor. Appl. Genet. **62:** 89–96.

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

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

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machine. J. Chem. Phys. **21:** 1087–1091.

Price, C. S. C., K. A. Dyer and J. A. Coyne, 1999 Sperm competition between Drosophila males involves both displacement and incapacitation. Nature **400:** 449–452.

Prout, T., and J. Bundgaard, 1977 Population genetics of sperm displacement. Genetics **85:** 95–124.

Communicating editor: G. Churchill