Properties of Maximum Likelihood Male Fertility Estimation in Plant Populations

M. T. Morgan

Department of Botany, Department of Genetics and Cell Biology, Washington State University, Pullman, Washington 99164-4238

Manuscript received November 18, 1997
Accepted for publication February 27, 1998

ABSTRACT

Computer simulations are used to evaluate maximum likelihood methods for inferring male fertility in plant populations. The maximum likelihood method can provide substantial power to characterize male fertilities at the population level. Results emphasize, however, the importance of adequate experimental design and evaluation of fertility estimates, as well as limitations to inference (*e.g.*, about the variance in male fertility or the correlation between fertility and phenotypic trait value) that can be reasonably drawn.

ONE half of the nuclear genes in most plants pass through the male reproductive pathway, yet estimates of male fertility based on ecological observations such as dispersal distances of pollen analogues or observed pollinator movements can be "disappointingly crude" (Snow and Lewis 1993, p. 332): any one of a large number of individuals capable of producing male gametes may potentially sire a particular offspring. This situation is attributable to unique features of plant biology, particularly the difficulty of reliably circumscribing the pool of potential fathers.

Genetic markers can assist male fertility estimation. The most powerful marker-based methods (Devlin *et al.* 1988; Roeder *et al.* 1989; Brown 1990; Adams *et al.* 1992) partition paternity among genetically possible fathers using a maximum likelihood argument (Roeder *et al.* 1989; Smouse and Meagher 1994). Estimated fertilities may be used to evaluate specific hypotheses (*e.g.*, that all males have equal fertility) and to describe patterns such as variation in male fertility (*e.g.*, Devlin and Ellstrand 1990; Devlin *et al.* 1992; Smouse and Meagher 1994) or the relationship between male trait value and fertility as a measure of selection (*e.g.*, Schoen and Stewart 1986; Broyles and Wyatt 1990; Conner *et al.* 1996).

Here I use computer simulation to document statistical power of maximum likelihood methods and to identify conditions when reasonable insight into male fertility variation can be obtained. The focus is on allozyme data, where factors contributing to manageable experimental designs are well understood; speculation on possible results from highly variable markers is presented in discussion. Results indicate the importance of genetic exclusion probability (ε, see Chakraborty *et al.* 1988; Devlin *et al.* 1988), number and size of maternal progeny arrays, and estimation of a limited number of fertilities. Future paternity studies require further mathematical analysis of maximum likelihood methods, or extensive computer simulation, to adequately evaluate the accuracy of inferences made.

Author e-mail: mmorgan@wsu.edu

MATERIALS AND METHODS

Maximum likelihood estimation: Smouse and Meagher (1994; following Roeder *et al.* 1989) develop a maximum likelihood estimator of male fertility for use in conjuction with electrophoretic or other genetic marker data. The problem is to estimate a vector λ of male fertilities, using a matrix **X** of genetic data. Each element of the fertility vector λ_j corresponds to the fertility of the *j*th unique male genotype, while the matrix entry X_{ij} is the probability of observing offspring genotype *i* given the genotypes of the maternal parent and the *j*th putative paternal parent (Devlin *et al.* 1988; Roeder *et al.* 1989). The likelihood of a vector of male fertilities, given observed offspring genotypes, is

$$L = \prod_i \left(\sum_j X_{ij} \lambda_j \right). \quad (1)$$

The goal is to identify the vector of male fertilities maximizing this likelihood.

A maximum of the likelihood can be found using the expectation maximization algorithm (Roeder *et al.* 1989, p. 373). One iteration of this algorithm transforms a value of male fertility λ_j to a value λ_j' using the formula

$$\lambda_j' = \sum_i \frac{X_{ij}\lambda_j}{\sum_l X_{il}\lambda_l}. \quad (2)$$

The product $X_{ij}\lambda_j$ in the numerator represents the expectation step, while the division and outer sum correspond to maximization. The algorithm used here starts with an initial vector of male fertilities λ in which elements are equal and sum to one, $\Sigma_j \lambda_j = 1$. Iteration proceeds until the change in the log of the likelihood is less than 10^{-5} per iteration.

Simulation methodology: Simulation was used to evaluate the statistical power of the estimation procedure and to evaluate inference about male fertility. Simulations centered around a "standard" parameter set. The standard set assumed a dioecious population of 25 male and 25 female parents, with 20 progeny assayed per maternal family. Genetic data in the standard set consist of eight loci, each with two equally frequent alleles (expected exclusion probability ε = 0.81; observed exclusions in simulations, *e.g.*, in Figure 1, are less than

this because of the finite number of paternal parents). This parameter set involves assaying a reasonable number of progeny for a combination of loci with exclusion probabilities toward the high end of that attainable with allozyme markers. Natural populations are likely to have more than 25 potential males, but the analyses presented below suggest that this realistic situation results in poor statistical properties. Loci are in Hardy-Weinberg and linkage equilibrium and are inherited in a Mendelian fashion. Parental genotypes are known without error. Expected male fertilities were chosen from a Gaussian distribution with mean equal to the number of progeny simulated and coefficient of variation equal to CV_g; zero fertility was assigned when negative deviates were drawn. The actual fertility coefficient of variation CV_m (*i.e.*, variation in male fertility realized in a simulation) includes this source of variation and an additional multinomial component associated with sampling. Numbers of male and female parents, progeny array size, and number of loci were varied one at a time, with CV_g ranging between zero and one (with $CV_g < 0.7$, virtually all males sire some offspring, whereas for $CV_g = 1$, the distribution of male fertilities is nearly Poisson and ~35% of males sire no offspring). Each parameter combination involved 500 replicates.

Statistical power was evaluated using the likelihood ratio statistic suggested by Roeder *et al.* (1989). The test asks whether estimated male fertilities significantly improve the likelihood of the data when compared with the initial equal fertility vector. The test subtracts the log of the likelihood in Equation 1 calculated with the estimated fertilities from the log of the likelihood with equal fertilities, and is symbolized as $\Delta \log L$. For each statistical test, 500 data sets were simulated assuming equal male fertility, $CV_g = 0$. The $\Delta \log L$ values from these simulations represent the null distribution against which fertility distributions with $CV_g > 0$ are to be compared. Statistical power for each scenario with $CV_g > 0$ is determined as the proportion of $\Delta \log L$ values more extreme (larger) than 95% of the values under the assumption of equal expected fertility.

Two measures were used to characterize estimated *vs.* actual fertilities. The first, $\widehat{CV}_m/CV_m$, compared the estimated to actual male fertility coefficient of variation (this is also the ratio of estimated and actual male fertility standard deviations because the mean estimated and actual male fertility is the same). The fertility coefficient of variation represents the opportunity for selection (Crow 1958; Arnold and Wade 1984, p. 710), and $\widehat{CV}_m/CV_m$ provides an indication of whether this opportunity will be over- or underestimated in paternity analyses. The second measure, ρ, is the correlation between estimated and actual fertilities. This correlation is important in analyses of selection attempting to correlate phenotypic trait value with a measure of fitness (Lande 1976; Lande and Arnold 1983) because ρ determines the maximum possible correlation between trait and fitness (Li 1955, p. 151). The variance of individual fertility estimates provides an important method of assessing accuracy (Roeder *et al.* 1989), but is not reported here because of its indirect relation to population fertility variation or selection analysis.

RESULTS

Simulation results in Figure 1 indicate that statistical power to reject the null hypothesis of equal male fertility can be high, provided that male fertility is not too uniformly distributed. Paternity analyses benefit from large progeny sizes, many maternal progeny arrays, many loci (high exclusion probabilities), and few paternal parents. The lower panels of Figure 1 suggest that the total number of progeny assayed is important because similar curves result when comparable total progeny are assayed (*e.g.*, 10 progeny from 25 mothers = 250 total progeny *vs.* 20 progeny from 12 mothers = 240 total progeny).

Estimation of the male fertility variance may be biased, and there may not be a strong correlation between actual and estimated fertility (Table 1). These difficulties are particularly apparent when the actual variance is limited or when many male fertilities are estimated. Even in scenarios with 12 loci and, hence, extraordinary exclusion probability (expected $\varepsilon = 0.92$), the maximum likelihood method overestimates variance in male fertility by 1.5- to 2-fold. With eight loci and moderate exclusion probability (expected $\varepsilon = 0.81$), the correlation between actual and estimated fertility ranges from 0.25, when there are many males with limited fertility variation, to 0.65, when substantial fertility variation among relatively few males is estimated using many or large maternal families. With the exclusion probability offered by 12 loci, the correlation between actual and estimated fertility can rise to 0.83. When males have equal expected fertility, replicates with 50 females or 40 progeny per female show a slight decrease in performance of the estimators compared with standard parameter values involving fewer females or progeny. A similar pattern is observed when male fertility variation is summarized as a ratio of expected values, rather than as the expected value of ratios, so that the difference is not likely to result from uncertainty in the denominator of $\widehat{CV}_m/CV_m$. Instead, this result may reflect an underlying bias in the imperfectly estimated fertilities, reinforced by larger sample sizes.

DISCUSSION

Maximum likelihood methods can detect significant male fertility variation when applied to appropriate data sets (Roeder *et al.* 1989). However, low statistical power (Figure 1), biased estimates of fertility variation, and low correlation between actual and estimated fertility (Table 1) occur with few loci, few maternal progeny arrays, few progeny per maternal family, or many potential fathers. The fertility coefficient of variation, and hence opportunity for selection, can be substantially overestimated, even with 12 loci and exclusion probability $\varepsilon = 0.92$. The correlation between estimated and actual fertility can reduce the correlation between trait value and relative fertility in a selection analysis by 50% or more (Table 1). These results suggest how experimental design can enhance statistical power, and they indicate limits to inference drawn from such experiments.

Experimental populations are well suited to inference of male fertility (Devlin and Ellstrand 1990; Devlin *et al.* 1992; Kohn and Barrett 1992; Conner *et al.* 1996), although some care must be taken in evaluating

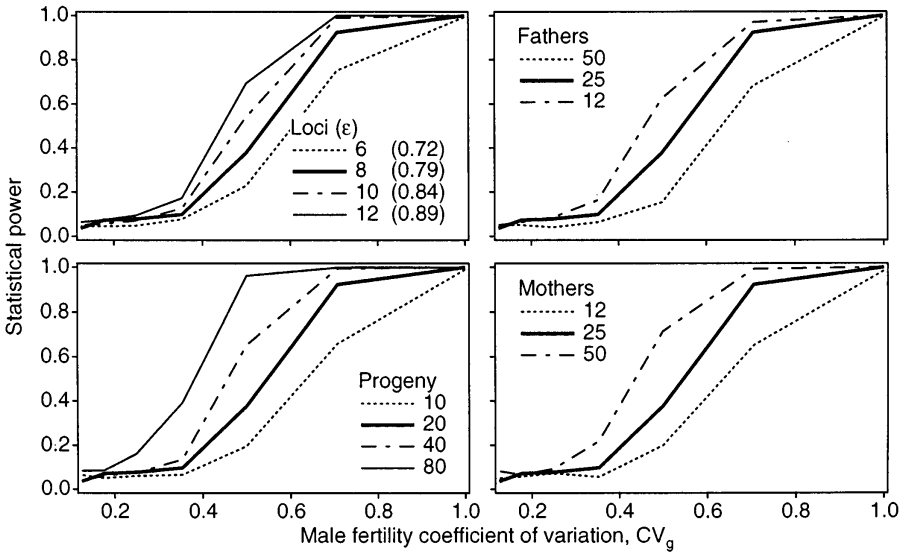

Figure 1.—Statistical power to reject the hypothesis of equal male fertility. Each panel shows the effect of one factor (number of loci with two equally frequent alleles, progeny array size, number of potential male parents, number of maternal progeny arrays) on power, when the Gaussian component of fertility variation, CV_g, is altered. The heavy, solid line in each panel represents standard parameter values (25 male and female parents, 20 progeny per female, eight loci with two equally frequent alleles). Observed exclusion probabilities for the standard parameters, but with different numbers of loci, are shown as ε in the upper left panel.

male fertility in natural populations. In experimental populations, the number of male fertilities requiring estimation can be small, and genotypes represented in the population can be chosen to ensure high exclusion probability. The most ambitious experimental study to date (Conner *et al.* 1996) involves 60 hermaphroditic plants, ~35 progeny per maternal parent, and exclusion probability between 0.85 and 0.89. Analysis by Conner *et al.* shows that the coefficient of variation of estimated individual male fertilities in this study is small (<5%). The results in Table 1 suggest that even in this data set, male fertility variation will be moderately overestimated, and the ability to detect selection on reproductive traits will be diminished by the imperfect correlation between estimated and actual fertility. Nonetheless, there is reasonable promise for application of paternity estimation techniques in populations of 25 possible paternal parents with substantial fertility and allozyme variation present. Clearly excluded as candidates for fertility estimation in nature are populations with large numbers of males (including species with extensive gene flow), populations with limited or moderate allozyme variation, or species with small progeny array sizes.

Genetic information (exclusion probability ε) plays a prominent but not exclusive role in male fertility estimation. For instance, all parameter sets involving eight loci in Figure 1 have the same exclusion probability, yet statistical power varies from near zero to one, depending

TABLE 1

Characterization of male fertility with allozyme markers

Scenario	$\widehat{CV}_m / CV_m$	ρ
Equal expected fertility, $CV_g = 0$		
Standard	6.92 (2.47–18.8)	0.40 (−0.01–0.72)
50 males and females	13.34 (6.60–28.1)	0.25 (−0.05–0.51)
50 males	10.02 (5.28–17.5)	0.26 (−0.06–0.54)
50 females	7.52 (2.45–20.6)	0.39 (0.01–0.70)
40 progeny	7.13 (2.23–16.1)	0.39 (−0.02–0.72)
12 loci	2.07 (1.00–3.96)	0.70 (0.40–0.90)
Substantial fertility variation, $CV_g = 0.5$		
Standard	3.40 (1.45–7.58)	0.56 (0.11–0.82)
50 males and females	6.33 (3.09–12.2)	0.36 (0.04–0.62)
50 males	6.29 (3.25–12.0)	0.34 (0.04–0.61)
50 females	2.75 (1.17–6.42)	0.65 (0.31–0.87)
40 progeny	2.68 (1.17–6.06)	0.63 (0.26–0.86)
12 loci	1.50 (0.91–2.57)	0.83 (0.59–0.95)

Estimated *vs.* actual male fertility coefficient of variation, $\widehat{CV}_m / CV_m$, and correlation between actual and estimated fertility, ρ. Each line in the table summarizes 500 replicates of the standard parameter set (25 male and female parents, 20 progeny per female, eight loci with two equally frequent alleles) or scenarios differing from the standard as indicated, when males have equal expected fertility ($CV_g = 0$) or substantial fertility variation ($CV_g = 0.5$). Numbers in parentheses represent the 95% confidence interval.

TABLE 2

Characterization of male fertility with highly polymorphic markers

Loci	Alleles	Number of potential male parents: 25	100	200
Estimated to actual male fertility coefficient of variation, $\widehat{CV}_m / CV_m$				
4	4	2.3 (1.05–4.59)	4.4 (2.96–6.15)	4.7 (3.46–6.25)
	6	1.2 (0.85–1.73)	1.9 (1.43–2.54)	2.4 (1.86–3.04)
	8	1.1 (0.89–1.30)	1.3 (1.07–1.60)	1.5 (1.29–1.88)
8	4	1.0 (0.88–1.25)	1.2 (1.01–1.44)	1.3 (1.13–1.57)
	6	1.0 (0.97–1.05)	1.0 (0.98–1.06)	1.0 (0.99–1.07)
	8	1.0 (0.99–1.01)	1.0 (0.99–1.02)	1.0 (0.99–1.03)
Correlation between actual and estimated fertility, ρ				
4	4	0.68 (0.39–0.89)	0.34 (0.11–0.54)	0.22 (0.08–0.37)
	6	0.91 (0.77–0.98)	0.69 (0.55–0.81)	0.52 (0.38–0.65)
	8	0.97 (0.91–0.99)	0.87 (0.79–0.93)	0.76 (0.67–0.84)
8	4	0.98 (0.95–1.00)	0.92 (0.87–0.96)	0.84 (0.77–0.90)
	6	1.00 (0.99–1.00)	1.00 (0.99–1.00)	0.99 (0.97–1.00)
	8	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (0.99–1.00)

Estimated *vs.* actual male fertility coefficient of variation, $\widehat{CV}_m / CV_m$, and correlation between actual and estimated fertility, ρ, with varying numbers of equally frequent alleles at four or eight loci. Each line in the table summarizes 500 replicates with 10 progeny assayed from 25 females (250 total progeny), with varying numbers of potential male parents having equal expected fertility ($CV_g = 0$). Numbers in parentheses represent the 95% confidence interval.

on other aspects of experimental design and the actual amount of fertility variation. The results of Table 1 similarly show the importance of factors other than exclusion probability in characterizing fertility variation. Even if exclusion were complete and fertility assigned without error, under the hypothesis of uniform expected male fertility, the error of individual fertility estimates follows a multinomial distribution with sampling variance inversely proportional to the total number of progeny surveyed (Roeder *et al.* 1989). Thus, the best strategy for increasing accuracy of fertility estimates may not be maximizing genetic exclusion (*e.g.*, through use of hypervariable markers). Perhaps the most encouraging result is the benefit of increasing the number of progeny sampled for statistical power (either sampling more progeny per mother or more maternal parents, see Figure 1) because assaying additional progeny is the factor most easily manipulated by the investigator interested in natural populations. Admittedly, Table 1 shows that increasing progeny sampled may only modestly increase the precision of estimated male fertility parameters.

Modern molecular markers may substantially expand the applicability of paternity analyses, although available data sets only hint at appropriate parameters for further investigation. Simple sequence repeats (SSRs) are one promising genetic marker with abundant polymorphism and codominant expression. Although many SSR loci are found in rice (Chen *et al.* 1997) or maize (Smith *et al.* 1997), published studies of natural plant populations document SSR variants at relatively few loci. For instance, four polymorphic loci with effective number of alleles (Hartl and Clark 1989, p. 126) between 1.9 and 5.24 were found in *Pithecellobium elegans* (Mimosoideae; Chase *et al.* 1996), while a single locus with six alleles was identified in the tropical tree *Gliricidia sepium* (Dawson *et al.* 1997). Table 2 shows simulation results when highly polymorphic loci are assayed in 250 progeny (10 offspring from 25 maternal parents) with between 25 and 200 potential male parents and male fertility differences resulting entirely from sampling (*i.e.*, $CV_g = 0$). Variation similar to that reported from natural populations (*e.g.*, four alleles at four loci) continues to provide biased estimates of male fertility variation and low correlation between actual and estimated fertility, even with only 25 potential male parents. A greater number of alleles per locus results in very favorable prospects for paternity analysis, but observation of many alleles per locus may be precluded by genetic drift in the small populations assumed here. Investing in development of additional loci offers very effective paternity analysis, even in moderate-sized populations.

Computer simulation and resampling techniques may continue to play an important part in paternity studies. Preliminary analysis, using knowledge of marker variation, population structure, and proposed experimental design, might help to determine whether a full-scale study will be informative (Roeder *et al.* 1989) and to identify an appropriate sampling strategy (*e.g.*, polymorphism such as that in Table 2 suggests few progeny per maternal parent compared with that in Table 1). Interpretation of hypothesis tests and inferences from a paternity study also requires investigation of statistical properties of the inference to determine the expected bias in estimates of male fertility variation or the ex-

pected correlation between estimated and actual fertility. Computer simulation also offers the opportunity to incorporate idiosyncrasies of the data set under investigation. For instance, using many marker loci increases the likelihood of linkage, parental genotypes may not be in Hardy-Weinberg proportions, and markers may violate Mendelian patterns of segregation.

Finally, the method of estimating paternity used here represents only one form of analysis. Adams and coworkers (Adams and Birkes 1991; Adams 1992; Burczyk *et al.* 1996) use electrophoretic data to estimate the fraction of self-fertilizations, matings between neighboring individuals, and mating between individuals outside the local neighborhood. Matings between neighboring individuals are further estimated as a function of plant or population attributes (*e.g.*, size of putative paternal parent, distance between maternal and putative paternal parent). This procedure has much to recommend it, because it restricts the pool of potential fathers (through estimation of neighborhood size) and directly estimates a small number of biologically interesting parameters (*e.g.*, relationship between plant size and fertility) rather than relying on intermediary estimates of a large number of male fertilities. These methods were developed for seed orchards with relatively few maternal parents and well-defined populations, so their application to natural populations should be approached with caution.

This research was supported by a Natural Sciences and Engineering Research Council of Canada postdoctoral fellowship. Daniel Schoen, Peter Smouse, and anonymous reviewers provided many helpful comments on earlier versions.

LITERATURE CITED

Adams, W. T., 1992 Gene dispersal within forest tree populations. New Forests **6:** 217–240.

Adams, W. T., and D. S. Birkes, 1991 Estimating mating patterns in forest tree populations, pp. 157–172 in *Biochemical Markers in the Population Genetics of Forest Trees*, edited by S. Fineschi, M. E. Malvolti, F. Cannata and H. H. Hattemer. SPB Academic Publishing, The Hague.

Adams, W. T., D. S. Birkes and V. J. Erickson, 1992 Using genetic markers to measure gene flow and pollen dispersal in forest tree seed orchards, pp. 37–61 in *Ecology and Evolution of Plant Reproduction*, edited by R. Wyatt. Chapman & Hall, New York.

Arnold, S. J., and M. J. Wade, 1984 On the measurement of natural and sexual selection: theory. Evolution **38:** 709–719.

Brown, A. H. D., 1990 Genetic characterization of plant mating systems, pp. 145–162 in *Plant Population Genetics, Breeding, and Genetic Resources*, edited by A. H. D. Brown, M. T. Clegg, A. L. Kahler and B. S. Weir. Sinauer Associates, Sunderland, MA.

Broyles, S. B., and R. Wyatt, 1990 Paternity analysis in a natural population of *Asclepias exaltata*: multiple paternity, functional gender, and the 'pollen-donation' hypothesis. Evolution **44:** 1454–1468.

Burczyk, J., W. T. Adams and J. Y. Shimizu, 1996 Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus attenuata* Lemmon.) stand. Heredity **77:** 251–260.

Chakraborty, R., P. E. Smouse and T. R. Meagher, 1988 Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genetics **118:** 527–536.

Chase, M., R. Kesseli and K. Bawa, 1996 Microsatellite markers for population and conservation genetics. Am. J. Bot. **83:** 51–57.

Chen, X., S. Temnykh, Y. Xu, Y. G. Cho and S. R. McCouch, 1997 Development of a microsatellite framework map providing genome-wide coverage in rice (*Oryza sativa* L.). Theor. Appl. Genet. **95:** 553–567.

Conner, J. K., S. Rush, S. Kercher and P. Jennetten, 1996 Measurements of natural selection on floral traits in wild radish (*Raphanus raphanistrum*). 2. Selection through lifetime male and total fitness. Evolution **50:** 1137–1146.

Crow, J. F., 1958 Some possibilities for measuring selection intensities in man. Hum. Biol. **30:** 1–13.

Dawson, I. K., R. Waugh, A. J. Simons and W. Powell, 1997 Simple sequence repeats provide a direct estimate of pollen-mediated gene dispersal in the tropical tree *Gliricidia sepium*. Mol. Ecol. **6:** 179–183.

Devlin, B., and N. C. Ellstrand, 1990 Male and female fertility variation in wild radish, a hermaphrodite. Am. Nat. **136:** 87–107.

Devlin, B., K. Roeder and N. C. Ellstrand, 1988 Fractional paternity assignment: theoretical development and comparison to other methods. Theor. Appl. Genet. **76:** 369–380.

Devlin, B., J. Clegg and N. C. Ellstrand, 1992 The effect of flower production on male reproductive success in wild radish populations. Evolution **46:** 1030–1042.

Hartl, D. L., and A. G. Clark, 1989 *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA.

Kohn, J. R., and S. C. H. Barrett, 1992 Experimental studies on the functional significance of heterostyly. Evolution **46:** 43–55.

Lande, R., 1976 Natural selection and random genetic drift in phenotypic evolution. Evolution **30:** 314–334.

Lande, R., and S. J. Arnold, 1983 The measurement of selection on correlated characters. Evolution **36:** 1210–1226.

Li, C. C., 1955 *Population Genetics*. The University of Chicago Press, Chicago.

Roeder, K., B. Devlin and B. G. Lindsay, 1989 Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. Biometrics **45:** 363–379.

Schoen, D. J., and S. C. Stewart, 1986 Variation in male reproductive investment and male reproductive success in white spruce. Evolution **40:** 1109–1120.

Smith, J. S. C., E. C. L. Chin, H. Shu, O. S. Smith, S. J. Wall *et al.*, 1997 An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays*): comparisons with data from RFLPS and pedigree. Theor. Appl. Genet. **95:** 163–173.

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

Snow, A. A., and P. O. Lewis, 1993 Reproductive traits and male fertility in plants—empirical approaches. Annu. Rev. Ecol. Syst. **24:** 331–351.

Communicating editor: A. H. D. Brown