

## Genetic Analysis of Male Reproductive Contributions in *Chamaelirium luteum* (L.) Gray (Liliaceae)

Peter E. Smouse\* and Thomas R. Meagher†

\*Center for Theoretical and Applied Genetics, and Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, New Jersey 08903-0231 and †Department of Biological Sciences, Rutgers University, Piscataway, New Jersey 08855-1059

Manuscript received February 15, 1993  
Accepted for publication September 23, 1993

### ABSTRACT

Genealogical analysis is a powerful tool for analysis of reproductive performance in both natural and captive populations, but assignment of paternity has always been a stumbling block for this sort of work. Statistical methods for determining paternity have undergone several phases of development, ranging from straightforward genetic exclusion to assignment of paternity based on genetic likelihood criteria. In the present study, we present a genetic likelihood-based iterative procedure for fractional allocation of paternity within a progeny pool and apply this method to a population of *Chamaelirium luteum*, a dioecious member of the Liliaceae. Results from this analysis clearly demonstrate that different males make unequal contributions to the overall progeny pool, with many males contributing essentially nothing to the next generation. Furthermore, the distribution of paternal success among males shows a highly significant departure from (Poisson) randomness. The results from the present analysis were compared with earlier results obtained from the same data set, using likelihood-based categorical paternity assignments. The general biological pattern revealed by the two analyses is the same, but the estimates of reproductive success are only modestly (though significantly) correlated. The iterative procedure makes more complete use of the data and generates a more sharply resolved distribution of male reproductive success.

THE genetic history and current structure of a population can be represented in terms of its genealogy (CANNINGS and THOMPSON 1981; CAVALLI-SFORZA and EDWARDS 1967; ELSTON and STEWART 1971). Genealogical analysis has become an increasingly important component of studies of mating behavior and gene dispersal in natural populations (SMITH and ADAMS 1983; ELLSTRAND and MARSHALL 1985; HAMRICK and SCHNABEL 1985; MEAGHER 1986; MEAGHER and THOMPSON 1987) and in the management of captive animal populations (MCCRACKEN and BRADBURY 1977; FOLTZ and HOOGLAND 1981; HANKEN and SHERMAN 1981). Early effort involved tracing mitochondrial and chloroplast markers through maternal lineages (AVISE, LANSMAN and SHADE 1979; CLEGG, RALSON and THOMAS 1984; PALMER, JORGENSEN and THOMPSON 1985); methods have more recently been developed that use nuclear genetic markers to study both maternal and paternal lineages (MEAGHER 1986; MEAGHER and THOMPSON 1986, 1987; DEVLIN, ROEDER and ELLSTRAND 1988; ROEDER, DEVLIN and LINDSAY 1989; DEVLIN, CLEGG and ELLSTRAND 1992).

These newer methods have their origin in human population studies, where the assignment of a mother to an offspring is frequently obvious on other than genetic criteria and parentage assessment amounts to a paternity determination. The statistical/genetic

methodology for detailed paternity analysis for human medicolegal work dates back 50 years (ESSEN-MÖLLER 1939) and in spite of a recent resurgence of debate (VALENTIN 1984; AICKIN 1984; LI and CHAKRAVARTI 1985; ELSTON 1986; THOMPSON 1986; MICKEY, GJERTSON and TERASAKI 1986) is well established. Such techniques have recently been applied to the analysis of natural plant populations (MEAGHER 1986), using a genealogical inference strategy first described by THOMPSON (1975, 1976a, 1976b, 1976c, 1986, 1987) and applied by MEAGHER (1986), MEAGHER and THOMPSON (1986) and THOMPSON and MEAGHER (1987).

The ultimate in paternity analysis would be to employ enough genetic markers to exclude all but one male from consideration for each offspring, enabling categorical assignment of paternity. That approach has been successfully applied in only a few cases (ELLSTRAND 1984; BROYLES and WYATT 1990). It is possible to compute the expected fraction of males excluded for a specified mother: offspring pair from standard Hardy-Weinberg and Mendelian assumptions, and we have recently extended that formulation to finite populations (CHAKRABORTY, MEAGHER and SMOUSE 1988). Even with modest departures from Hardy-Weinberg equilibrium, there is a regular relationship between the allele-frequency profiles and the expected exclusion probability. Adding informative

loci always helps, but even with an extensive battery of genetic markers, a substantial fraction of all births yield equivocal paternity determinations. We have shown theoretically (CHAKRABORTY, MEAGHER and SMOUSE 1988) that the number of markers needed to assign paternity categorically for every offspring, based on exclusion criteria, is beyond the scope of allozyme analysis. Experience with organisms as diverse as plants (ELLSTRAND and MARSHALL 1985; HAMRICK and SCHNABEL 1985; MEAGHER 1986), rodents (FOLTZ and HOOGLAND 1981; HANKEN and SHERMAN 1981), hymenopterans (HUGHES and QUELLER 1993), and bats (MCCRACKEN and BRADBURY 1977) shows this to be general.

Various workers have begun to augment the available polymorphisms with some of the newer DNA techniques, using single locus RFLPs or RAPDs (BAIRD *et al.* 1986; HEDRICK 1992; WILLIAMS *et al.* 1990), single locus VNTR markers (NAKAMURA *et al.* 1987), or the genomic minisatellite markers (JEFFREYS, WILSON and THEIN 1985a, 1985b; JEFFREYS *et al.* 1986; BURKE and BRUFORD 1987; JEFFREYS and MORTON 1987; QUINN and WHITE 1987; WETTON *et al.* 1987; HUGHES and QUELLER 1993). We have examined the utility of using an RFLP battery (SMOUSE and CHAKRABORTY 1986), showing that by using more than one restriction site per probed segment, we can convert a two-allele to a multiple-allele system. Multiple-allele systems are more information rich than are two-allele systems but are of limited applicability, due to linkage phase ambiguity of multimer heterozygotes. Multilocus minisatellite probes show promise as markers for small sets of candidate males, but there are considerable difficulties with genetic interpretation (HILL 1987; LYNCH 1988). Single-locus VNTR probes show promise, although there remain some lingering difficulties with phenotypic overlap of length alleles that require special attention (*e.g.*, BUDWOLE *et al.* 1991; DEVLIN, RISCH and ROEDER 1992). The recent introduction of RAPDs and the ability to generate large numbers of markers has reopened the question of exclusion-based categorical assignment, but the dominance exhibited by these loci reduces their resolving power (LEWIS and SNOW 1992; MILLIGAN and MCMURRY 1993). The major requirements for parentage analysis are that the markers should be unambiguously inherited, segregate independently in the population and lead to lower levels of ambiguity than the parentage uncertainty we are trying to resolve; current DNA methods do not (yet) meet that standard.

While a strictly exclusionary solution to the problem of parentage assessment is not yet generally attainable, we can nevertheless obtain partial resolution with only a modest number of polymorphic markers if we relax the criterion of success, assigning paternity on the

basis of genetic likelihood of being the male parent (THOMPSON 1975, 1976a, 1976b; MEAGHER 1986, 1991; THOMPSON and MEAGHER 1986; MEAGHER and THOMPSON 1987). This approach has two limitations. First, categorical assignments cannot be made for all offspring, due either to redundancies among male genotypes or to a likelihood profile for particular progeny that is largely ambiguous. We are unable to use all of the data. Second, because one of the homozygotes will always give a higher likelihood score than a heterozygote for a given genetic locus, all offspring with the allele will be allocated to homozygous males (DEVLIN, ROEDER and ELLSTRAND 1988; ADAMS, GRIFFIN and MORAN 1992). A homozygote does have a higher likelihood of being the transmitter of a given allele than does a heterozygote, so assignment to the homozygote is proper, but the net effect is a statistical bias in favor of homozygotes. The size of the bias decreases as we add genetic markers to the battery.

The next stage in methodological development has been to evaluate the distribution of paternity by making fractional paternity assignments of offspring genotypes, where each male's fraction is based on his relative likelihood of paternity (SCHOEN and STEWART 1986, 1987; DEVLIN, ROEDER and ELLSTRAND 1988; ROEDER, DEVLIN and LINDSAY 1989). This approach is applicable when the population distribution of male reproductive success is the analytical goal. This approach makes more complete use of the full data set and leads to genetically unbiased estimates of relative male reproductive contributions.

We have four objectives. We (1) describe a likelihood-based, iterative fractional allocation approach to estimate the relative reproductive successes of a collection of males, and (2) develop both likelihood ratio and nonparametric testing procedures to assess whether males contribute unequally and whether some of them can be ignored altogether. On the biological side, we (3) examine male reproductive contributions in a population of *Chamaelirium luteum* (L.) Gray (Liliaceae), a dioecious species whose population biology has been studied in detail (MEAGHER 1982; MEAGHER and ANTONOVICS 1982), and (4) compare the results obtained from the likelihood-based iterative fractional allocation treatment with those obtained from the earlier likelihood-based categorical assignment procedure (MEAGHER 1986, 1991).

#### THE ANALYTIC FRAMEWORK

The distribution of parentage in natural populations is generally uneven, a fact that is of interest in a number of contexts; we need a formulation designed to address that unevenness. In many cases, parentage inference is best done with genetic evidence. Maternity can generally be established unambiguously, reducing our problem to the estimation and testing of

unequal male reproductive contributions.

Consider a set of  $N$  mother:child pairs ( $MC_i$ ;  $i = 1, \dots, N$ ) and a set of  $K$  potential fathers ( $F_k$ ;  $k = 1, \dots, K$ ). We assume here that each child of a particular mother is an independent draw from the distribution of fathers. For none of the offspring do we know paternity, but we do know the situation can be represented as in Table 1a. One of the elements in the  $i$ th row is unity and all the others are zero; the row totals ( $N_{i\cdot}$ s) are all unity. We are interested in estimating the relative male reproductive contributions. Using the column totals, we define the array of relative male reproductive contributions (the frequency spectrum)  $\Lambda = \{\lambda_k\}$  as

$$\lambda_k = N_{\cdot k}/N \quad \text{for } k = 1, \dots, K. \quad (1)$$

The column totals of interest, the  $N_{\cdot k}$  are unknown. In the absence of genetic information, the probability that the  $k$ th male has fathered the  $i$ th offspring is

$$\lambda_k^{(0)} = 1/K \quad \text{for } k = 1, \dots, K. \quad (2)$$

This is the usual "no information" (equal contribution) solution. It represents the null hypothesis and a point of departure; in general, we suspect it is not valid.

We need some additional (preferably genetic) data if we are to reject this equal contribution solution in favor of an unequal contribution alternative. We can deal with virtually any inheritance pattern, including dominance and linkage, as long as the Mendelian segregation patterns of the markers are known (cf. SMOUSE and ADAMS 1983; SMOUSE and CHAKRABORTY 1986). Given the genotypes of a mother ( $M_i$ ) and a putative father ( $F_k$ ), the probability of obtaining the genotype of the offspring ( $C_i$ ) is

$$X_{ik} = Pr(\text{genotype of } C_i | \text{genotypes of } M_i \text{ and } F_k), \quad (3)$$

numbers that depend on Mendelian ratios and, if there is dominance or linkage, population gametic frequencies (SMOUSE and ADAMS 1983). Every cell of Table 1 has an attached value of  $X_{ik}$ ; the  $X_{ik}$  are constant for all that follows, and a genetic exclusion implies that  $X_{ik} = 0$ . Exclusions are useful, but nothing that follows requires overt exclusions.

Under random mating, the likelihood of obtaining the observed array of offspring ( $C_i$ ;  $i = 1, \dots, N$ ) from the known mothers ( $M_i$ ;  $i = 1, \dots, N$ ), and given the array of potential fathers ( $F_k$ ;  $k = 1, \dots, K$ ), is

$$L(\Lambda | \text{genetic data}) = \prod_{i=1}^N \Delta_i \\ = \prod_{i=1}^N [\lambda_1 X_{i1} + \lambda_2 X_{i2} + \dots + \lambda_K X_{iK}]. \quad (4)$$

In the absence of genetic data, we initially assume that the  $\lambda$ s are all the same, but we have every reason to

TABLE 1

Data structure for likelihood-based paternity analysis

Mother	Candidate fathers				Totals
	$F_1$	$F_2$	—	$F_K$	
$M_1$	${}^a N_{11}$	$N_{12}$	—	$N_{1K}$	1
$M_2$	$N_{21}$	$N_{22}$	—	$N_{2K}$	1
—	—	—	—	—	—
$M_N$	$N_{N1}$	$N_{N2}$	—	$N_{NK}$	1
Totals	${}^b N_{\cdot 1}$	$N_{\cdot 2}$	—	$N_{\cdot K}$	$N$
Mother and offspring	Putative fathers				Totals
	$F_1$	$F_2$	—	$F_K$	
$MC_1$	${}^c P_{11}$	$P_{12}$	—	$P_{1K}$	1
$MC_2$	$P_{21}$	$P_{22}$	—	$P_{2K}$	1
—	—	—	—	—	—
$MC_N$	$P_{N1}$	$P_{N2}$	—	$P_{NK}$	1
Averages	${}^d \lambda_1$	$\lambda_2$	—	$\lambda_K$	1

<sup>a</sup> Numbers of offspring ( $N_{ik}$ ) of known mothers ( $M_i$ ) and putative fathers ( $F_k$ ).

<sup>b</sup> The individual male reproductive contributions ( $N_{\cdot k}$ s) are unknown.

<sup>c</sup> Probabilities of paternity ( $P_{ik}$ ) for each of the potential fathers ( $F_k$ ), given known mother/child pairs ( $MC_i$ ).

<sup>d</sup> Maximum likelihood estimates of the reproductive contributions ( $\lambda_k$ ) of males.

expect uneven male reproductive contributions. The essence of the analysis is that we need to maximize  $L$ , relative to the choices of unequal  $\lambda$ s. There are several ways to obtain maximum likelihood estimates, all of them iterative and none of them trivial (ROEDER, DEVLIN and LINDSAY 1989; DEVLIN, RISCH and ROEDER 1992), but having arrived at the solutions, we can replace the  $N_{ik}$ s in Table 1 with a set of  $P_{ik}^{(*)}$ s

$$P_{ik}^{(*)} = \lambda_k^{(*)} X_{ik} / \Delta_i^{(*)}. \quad (5)$$

The results are presented as Table 1b. Each row sums to unity and the table contains an explicit zero ( $P_{ik}^{(*)} = 0$ ) wherever there is a genetic exclusion ( $X_{ik} = 0$ ). The relative column totals are the maximum likelihood estimates of male reproductive contributions

$$\lambda_k^{(*)} = \sum_{i=1}^N P_{ik}^{(*)} / N \dots \quad (6)$$

We have employed an expectation maximization (EM) algorithm to obtain an iterative solution, a method first described by DEMPSTER, LAIRD and RUBIN (1977) and more recently applied to this problem by ROEDER, DEVLIN and LINDSAY (1989). Begin with the uniform spectrum in Equation 2. Then determine an initial set of estimates

$$P_{ik}^{(0)} = \lambda_k^{(0)} X_{ik} / \Delta_i^{(0)}. \quad (7)$$

With these estimates of the  $P_{ik}$ s, obtain new estimates of the  $\lambda$ s

$$\lambda_k^{(1)} = \sum_{i=1}^N P_{ik}^{(0)} / N \dots \quad (8)$$

We have now arrived at the final solution obtained in the early methods presented for fractional paternity analysis (DEVLIN, ROEDER and ELLSTRAND 1988). These estimates use the same Mendelian and Hardy-Weinberg underpinnings as the estimates derived from likelihood-based categorical allocation (MEAGHER 1986) but are still not the full maximum likelihood estimates.

There is no real need to accept even these first-pass fractional estimates, however, and we can use the  $\lambda_k^{(l)}$ -estimates to obtain updated estimates of the  $P_{ik}$

$$P_k^{(l)} = \lambda_k^{(l)} X_{ik} / \Delta_i^{(l)}. \tag{9}$$

To obtain the maximum likelihood estimates, ROEDER, DEVLIN, and LINDSAY (1989) simply iterated the  $\lambda_k$ s and  $P_{ik}$ s to convergence. The algorithm is guaranteed to yield at least a local maximum. If the  $\mathbf{X}$ -matrix is of full rank (tantamount to the statement that the effect of each male is separately identifiable), we can expect a global maximum. In the event that the  $\mathbf{X}$ -matrix is not of full rank (with some male genotypes represented more than once in the data set or other linear dependencies in the progeny array), multiple solutions are possible (DEVLIN, ROEDER and ELLSTRAND 1988).

Given a set of  $\lambda_k^{(*)}$  that are not all equal, we need a test of whether the observed departure from uniformity is statistically convincing. The null hypothesis of male reproductive uniformity is represented by Equation 2, which yields a likelihood value of

$$L(H_0|X) = \prod_{i=1}^N \Delta_i^{(0)} = \prod_{i=1}^N [\lambda_1^{(0)} X_{i1} + \lambda_2^{(0)} X_{i2} + \dots + \lambda_K^{(0)} X_{iK}], \tag{10}$$

where  $\mathbf{X} = \{X_{ik}\}$  is the matrix of probabilities described by Equation 3. The alternative (unequal contribution) hypothesis yields a likelihood given by

$$L(H_a|X) = \prod_{i=1}^N \Delta_i^{(*)} = \prod_{i=1}^N [\lambda_1^{(*)} X_{i1} + \lambda_2^{(*)} X_{i2} + \dots + \lambda_K^{(*)} X_{iK}]. \tag{11}$$

The usual test of divergence from the null hypothesis is

$$\chi_{0a}^2 = -2[\log L(H_0|X) - \log L(H_a|X)], \tag{12}$$

asymptotically distributed as  $\chi^2$  with  $K - 1$  degrees of freedom under the null hypothesis of reproductive equality, where  $K$  is the dimension of the  $\mathbf{X}$ -matrix.

MALE REPRODUCTION IN *C. luteum*

**The genetic data:** We have used the iterative EM algorithm to analyze reproductive contributions of

TABLE 2

Allozyme loci used to provide parentage resolution in *C. luteum*

Genetic locus	No. of alleles	Exclusion probability	Cumulative probability
PGI	4	0.347	0.347
MPI	5	0.287	0.534
PGM	3	0.215	0.634
GDH	2	0.112	0.675
TPI-3	3	0.082	0.702
GOT-2	2	0.071	0.723
GOT-3	2	0.009	0.726
TPI-2	2	0.007	0.728

	No. of different genotypes	No. of male replicates
	1	16 <sup>a</sup>
	2	9
	1	8
	1	7
	1	6
	1	5
	4	4
	13	3
	28	2
	102	1

<sup>a</sup> Numbers of replicate 8-locus genotypes among the 273 candidate males.

males from a natural population of *C. luteum* in Orange County, North Carolina, a population that has been the object of long-term demographic and genealogical study (MEAGHER 1982, 1986, 1991; MEAGHER and ANTONOVICS 1982; MEAGHER and THOMPSON 1986, 1987; CHAKRABORTY, MEAGHER and SMOUSE 1988; THOMPSON and MEAGHER 1987). This species is a long-lived dioecious forest-floor perennial that shows a strong degree of sexual dimorphism and that typically exhibits male-biased sex ratios. These genetic data on 2255 offspring of 70 known mothers and 273 candidate males represent the breeding population within the Natural Area site in 1981. TRM has characterized all of these individuals for a set of eight polymorphic loci (PGI, PGM, GOT2, GOT3, TPI2, TPI3, GDH and MPI). The numbers of alleles per locus and expected exclusion probabilities are presented in Table 2a. There are 154 different 8-locus genotypes among the 273 males, with the replicate numbers as indicated in Table 2b. As we add loci in order of their decreasing contributions to exclusion probability (parentage resolution), diminishing returns set in quickly. Recall that the exclusion probability for a multiple-locus battery is

$$E(L \text{ loci}) = 1 - \prod_{l=1}^L (1 - E_l), \tag{13}$$

where  $E_l$  is the exclusion probability for the  $l$ th locus. The overall exclusion probability is ~73%, but the last three loci contribute almost nothing to resolution. Even with considerable genetic information, only 520

of the progeny can be assigned a (relatively) unambiguous father on the basis of likelihood criteria, and only 55 a categorically unambiguous father on the basis of exclusion criteria. This is precisely what we should expect on theoretical grounds (CHAKRABORTY, MEAGHER and SMOUSE 1988). Such partial resolution has sometimes been viewed as precluding useful inference on male reproductive contributions, but we will show that is very far from the case.

There are 615,615 cells of the **X**- and **P**-matrices, and we can impose over 543,000 zeroes due to genetic exclusions, reducing the number of nonexcluded fathers for the average offspring from 273 to 33. Moreover, the Mendelian probabilities ( $X_{ik}$ -values) for the unassigned progeny, given particular candidate fathers, range from 1 to  $(1/2)^{16}$ , which contributes to parentage resolution. The Mendelian profiles among these unassigned progeny are so uneven that the 575 assignments mentioned above are almost categorically resolvable on likelihood criteria (MEAGHER 1986). As noted above, this likelihood-based allocation strategy has drawn some criticism (DEVLIN, ROEDER and ELLSTRAND 1988; ADAMS, GRIFFIN and MORAN 1992) because it does not use all the progeny data, so precision is suboptimal, and because it yields genetically biased assignments. As exclusion probability increases, precision improves and the bias decreases. The iterated proportional allocation procedure described here has no genetic biases, but the impact of minor contributors is slightly overestimated at the expense of the major contributors, whatever their respective genotypes.

**Analysis of male reproductive contributions:** We subjected the full **X**-matrix to the likelihood analysis described above, using a program called CHAM-LAMB (available from P.E.S. on request), and we obtained a very uneven array of estimated male contributions. For male genotypes with multiple replicates, there is no way to apportion their collective contribution meaningfully among individuals, but the collective contribution is well estimated; we have 154 different male genotypes and 153 degrees of freedom. Even accounting for the fact that some male genotypes were represented several times in the population (Table 2b), there is no apparent relationship between population frequency of male genotypes and their reproductive contributions. We computed the log-likelihood ratio test criterion, and from it extracted an approximate  $\chi^2$  test criterion ( $\chi^2 = 411.20$ , 153 d.f.) via Equation 12. There is compelling evidence that parentage is uneven in this population.

It is also of interest to determine how many males we can delete from the candidate list with virtually no loss of information. We can compare the log-likelihood value obtained with the full set of males against that obtained with a reduced set, but the test criterion

is no longer  $\chi^2$  distributed, because the subhypothesis is a side solution. Whether the test is approximately  $\chi^2$  distributed or not, the results of reducing the number of candidate males are nevertheless revealing (Table 3); we can clearly dispense with as many as 65–83 males with little or no loss of information. Given that a large number of males contribute nothing, is it possible that the variance in reproductive contribution is solely due to the fact that some males do not contribute at all, while the rest contribute equally? The ratio of largest to smallest contribution among the remaining males is 286:1; the distribution of male reproductive contributions among the real contributors appears to be uneven. To determine whether contributions among the remaining males are significantly uneven, we computed a log-likelihood test of the null hypothesis that the contributing males all had equal contributions, with the noncontributors assigned  $\lambda$ -values of zero. These results are shown in the last column of Table 3, and clearly show that the equal contribution hypothesis for a smaller subset of males is untenable. Many males (65–83) contribute nothing, and reproductive contributions among the rest are extremely uneven.

**A nonparametric alternative:** The results are unequivocal, but we need a more visual way of gauging departures from equal reproductive contributions. Suppose we order the males in terms of their respective contributions, largest contributors first and smallest last. We then add males to the battery in decreasing order of their contributions, tallying the cumulative fraction of parentage accounted for as we go along. We then plot the results, as we have done for our *C. luteum* example in Figure 1a. The 45° line is what we should expect from an exactly even distribution of parentage; each additional male has exactly 2255/273 progeny, a perfect fit to the null hypothesis. The shaded area between the cumulative curve and the 45° line is an increasing measure of the unevenness of parental contributions. Even under the null hypothesis of equal fitness, of course, we expect some variation in the numbers of offspring sired by different males, due to reproductive sampling. The curve will depart somewhat from the 45° line even if the parametric contributions are uniform; we expect to see the sort of empirical result shown in Figure 1b. Any measure of departure from the null hypothesis should allow for the difference in the areas in Figure 1a and Figure 1b. We should even expect some of the males to leave no progeny, just by chance. We need a null distribution for the result in Figure 1b, against which to compare the result in Figure 1a. The difficulty is that even with the best maximization algorithms now available (see ROEDER, DEVLIN and LINDSAY 1989; DEVLIN, RISCH and ROEDER 1992), computational speed is so slow as to be prohibitive for the

TABLE 3

Numbers of males not contributing and the information content of their removal, measured as  $-2[\log L(H_A) - \log L(H_0)]$ , where  $H_A$  is an intermediate hypothesis of a subset of males contributing unequally and the rest contributing nothing, or measured as  $-2[\log L(H_B) - \log L(H_0)]$ , where  $H_B$  is an intermediate hypothesis of a subset of males contributing equally and the rest contributing nothing

No. of noncontributors		Unequal Contributions $-2[\log L(H_A) - \log L(H_0)]$		Equal contributions $-2[\log L(H_B) - \log L(H_0)]$	
Genotypes	Individuals	Incremental	Cumulative	d.f.	$\chi^2$
51	66	0.00	0.00	102	380.87
57	75	3.29	3.29	96	388.19
65	83	2.70	5.99	88	408.13
80	111	46.70	52.69	73	377.15
123	231	1016.78	1069.47	30	817.69

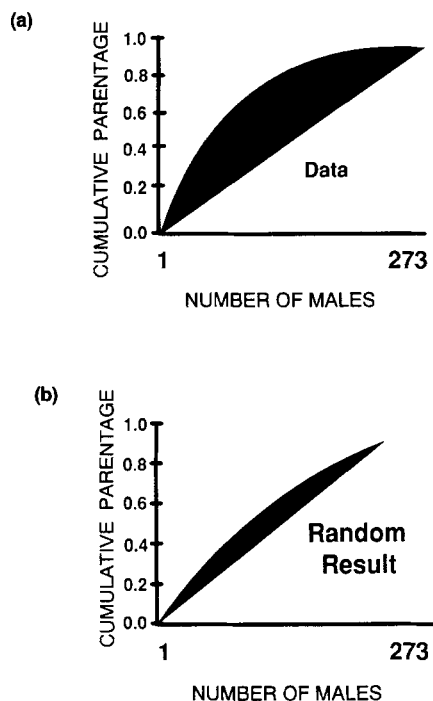


FIGURE 1.—Cumulative fraction of parentage accounted for by adding successive males to the battery in rank order of their contribution, best first. The straight diagonal represents a null hypothesis in which each male contributes equally, with the shaded area representing the degree of parentage unevenness. The two panels represent (a) the realized distribution from genetic analysis of *C. luteum* and (b) average distribution from random allocation of 2255 progeny to 273 males.

necessarily large scale simulation study required. We need something simpler and faster.

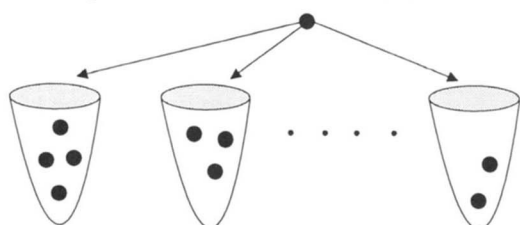
We can also formulate the allocation of offspring to fathers as an urn problem, with 273 urns and 2255 balls. Each father produces progeny according to a random (Poisson) distribution, with all fathers having the same (null hypothesis) expected number of offspring. The multivariate distribution of these 273 Poisson processes, subject to the restriction that there are  $n = 2255$  progeny, is multinomial with all urns having the same probability, specifically  $\lambda_k^{(0)} = 1/273$ . We randomly allocate each of the 2255 balls to an urn, with the probability of allocating any particular

ball to any particular urn being the same for all balls and urns, specifically  $\lambda_k^{(0)} = 1/273$ . This is tantamount to assigning each individual (ball) unambiguously to a single father (urn) and avoids the computational complexity of the full mixture analysis, where membership is less than categorical. Having randomly allocated the 2255 balls, we construct our cumulative fraction of paternity curve, as in Figure 1b and compute the area under it; we also tally the number of empty urns. We repeat the whole experiment a large number of times (say 10,000), and construct a null distribution of those criteria. The mean and variance of the number of empty urns is computable in closed form from the theory (see CHAKRABORTY 1993), but the area of the sector is most easily obtained by enumeration; in fact, the curve in Figure 1b is the average for 10,000 random runs. We present a comparison of the observed and random areas from 10,000 random urn samples in Figure 2, each area presented as its square root.

It is important to point out here that this is a *conservative* test; the likelihood estimates of the respective contributions from the mixture distribution are slightly biased upward for the minor contributors and downward for major contributors, thus flattening the curve (Figure 1a), relative to what we would have if the identity of each father were known without error, as is the case with the urn sampling (Figure 1b). The test is “stacked” in favor of the null hypothesis, and the departure of our *Chamaelirium* results from random expectation shown in Figure 2 is thus an *underestimate*. As was also evident from Table 3, the unevenness of male reproductive contributions is overwhelmingly established but here the result is visually obvious.

The results for empty urns are even more compelling; in 10,000 random runs, there were 28 runs with two empty urns and 755 runs with one empty urn; all the rest had no empty urns. For the actual results, we had 65–83 males contributing nothing. In keeping with the spirit of Table 3, is it possible that by simply discarding some males, the distribution of the areal

## Randomly Allocate 2255 Balls to 273 Urns



Compute Cumulative Parentage,  
Tally and Repeat 10,000 Times

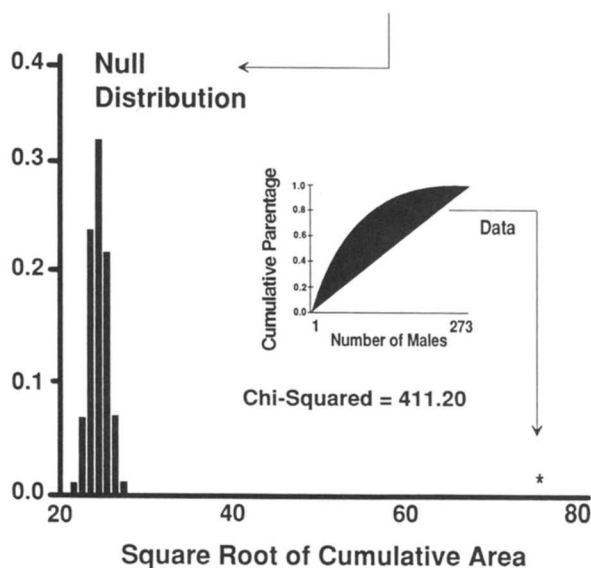
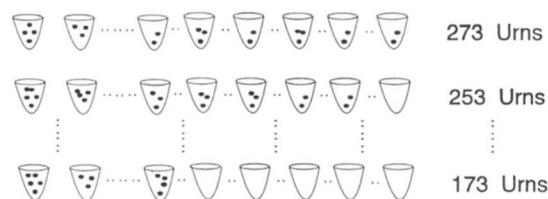


FIGURE 2.—A comparison of the areal measure of unevenness (Figure 1) from a genetic analysis of *C. luteum* and the distribution of areal measures drawn from 10,000 random allocations of 2,255 progeny to 273 males.

criterion in Figure 1b would approach that of Figure 1a more closely than indicated by Figure 2? We present the urn sampling equivalent of Table 3b in Figure 3. By randomly assigning balls to a subset of urns, we construct a *generous* null distribution for the equal contribution hypothesis with a smaller number of contributing males. The more males we exclude from consideration, the further to the right the null distribution moves, but it does not approach the data outcome for even 100 excluded males. The test is *conservative*, biased in favor of the null hypothesis. We cannot account for the nonuniformity of male reproductive contributions by excluding a large number of males and assuming uniformity of the rest!

**A comparison of methods:** The final estimates of  $\lambda$  obtained in the present study are directly analogous to the relative numbers of progeny assigned to each male in the likelihood-based categorical assignment procedure applied previously to these same data (MEAGHER 1986). A comparison of the two sets of estimates involves two considerations: (1) a comparison of the overall frequency spectra of male parentage (Figure 4a) and (2) a male-by-male comparison

## Randomly Allocate 2255 Balls to:



Tally Cumulative Parentage  
Repeat 10,000 Times  
Generate Null Distributions

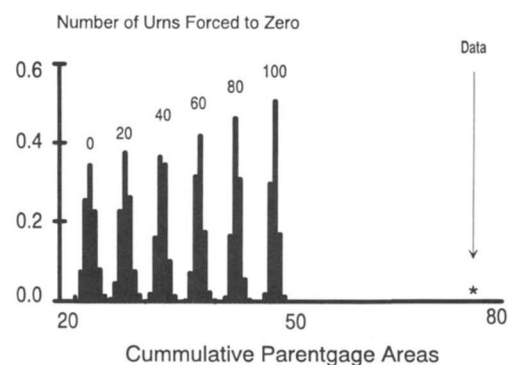


FIGURE 3.—A comparison of the areal measurement of unevenness (Figure 1) from a genetic analysis of *C. luteum* and the distribution of areal measures drawn from 10,000 random allocations of 2,255 progeny to subsets of the 273 males, with 20–100 males excluded, and with the null-distribution being that shown in Figure 2.

of estimated reproductive success (Figure 4b). The likelihood-based categorical assignment procedure (MEAGHER 1986) was restricted to a subset of 102 males with unique genotypes, to which 575 offspring could be assigned; we therefore used the same subset of males from the current analysis. If the 102 males are arrayed in order of increasing estimated reproductive success by each of the two methods, and if these ordered arrays of success are plotted against each other, we obtain the results in Figure 4a. The double-shouldered curve reflects the fact that in the earlier study (MEAGHER 1986) males were assigned an integer number of offspring (of the 575 that were assignable), while here they were assigned fractional proportions of 2255 offspring, the sum total of which was not generally integer valued. “Clumpiness” of the first distribution does not detract from the overall correlation ( $r = 0.95$ ). The overall message from the present analysis is the same as that from the earlier treatment; male parentage is highly uneven. The ordering of the two sets of estimates might not be the same, of course, so we present a male-by-male comparison of relative contributions ( $\lambda$ s) in Figure 4b. The correlation is modest at best ( $r = 0.32$ ), though significant, showing that the two procedures yield differences in detail. Given that the maximum likeli-

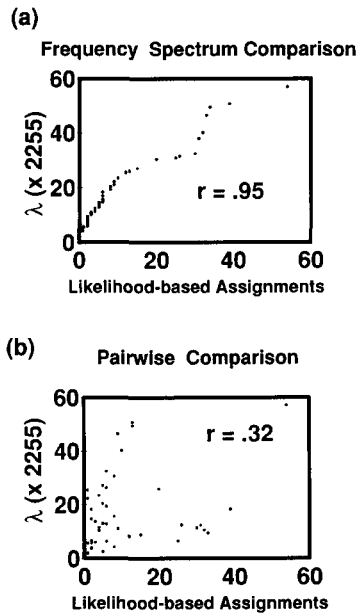


FIGURE 4.—The relationship between male parentage distributions under likelihood-based categorical assignment and likelihood-based iterative fractional allocation: (a) overall frequency spectra for 102 genetically unique males, presented in rank order of performance for the two methods; (b) male-by-male comparison, plotted as pairs of  $\lambda$ -values under the two treatments.

hood procedure uses all of the offspring and all of the males, the current estimates are much to be preferred.

#### DISCUSSION

The ability to measure differences in reproductive output for both captive and feral populations is central to a number of important problems in population biology and breeding colony management. The use of genetic markers to infer paternity has been gaining popularity as a means of determining male reproductive success. One of our primary interests is the study of factors influencing reproductive success in natural populations. Our goal in developing the method presented here is to utilize genetic information within a population to measure directly the properties of interest, without the necessity of categorically identifying specific parents. Our emphasis is less on identification than it is on estimation and hypothesis testing. Methodological controversy that has arisen in studies of paternity, both in natural populations and in forensic applications, has resulted in a shift of attention from the basic biological issues underlying the exercise onto identification procedures *per se*. By focusing directly on the end product, the biological hypothesis being tested, the present approach should help return attention to the original biological motivation for the analysis.

The maximum likelihood approach used here, developed originally by ROEDER, DEVLIN and LINDSAY (1989), makes assumptions about the mating biology of the population. We have explicitly assumed that

the male parentage for each seedling is an independent draw from the male parent (pollen) pool, that  $\lambda$  is solely a statement about overall male effectiveness. Specifically, we have assumed that for a given male,  $\lambda$  is the same for all females (no pairwise fertility effects) and that each seedling from a given female is an independent draw from the pollen pool (no tendency for clustering of male parentage by female). In gauging the overall impact of unequal male parentage, this is not an unreasonable first approximation assumption, but on biological grounds, it is almost surely an oversimplified model of the mating system.

In particular, *C. luteum* is an entomophilous species, and it would not be surprising if the multiple ovules of a single mother showed some clustering of male fertilization. The maternal plant serves as a common location for pollen delivery, and insect-vectored pollen is rarely delivered in single grain units. It is possible that pollination events, rather than single pollen grains, are the independent events of interest. For the present analysis, we have not taken advantage of the maternal sibship structure of the data set, but we have similar analyses in train that are designed to explore that alternative view of the breeding biology. The conclusions from the present analysis of *C. luteum* are clear-cut and compelling. Male reproductive success is unevenly distributed, with many males contributing little (if anything) to the progeny pool. The variance in male reproductive success, an important component of current models of breeding system evolution in plants (CHARLESWORTH, SCHEMSKY and SORK 1987; LYONS *et al.* 1989), is obviously large, suggesting strong opportunities for natural selection. In previous analyses using the likelihood-based categorical assignment procedure, little relationship was found between male morphological features and reproductive success. On the other hand, spatial proximity to other males did influence reproductive success, possibly due to local competition for pollinator service (MEAGHER 1991). We are in the process of extending the method of the present paper to an analysis of male feature profiles and spatial aspects of the mating system in *C. luteum*. The maximum likelihood methods used here can be generalized to a wider class of problems of interest to the population biologist. The larger question is not so much "Which males produce which progeny?" as it is "Why do some males contribute disproportionate numbers of offspring?" Likelihood methods, focusing on estimation and hypothesis testing, should move us along the path toward addressing that larger and far more interesting question.

We thank C. KOBAK for help with the computer algorithms, computations and computer artwork. We thank D. COSTICH, B. DEVLIN, E. ELLE, O. GAGGIOTTI, C. KOBAK, L. REINERTSEN MEAGHER and D. SCHOEN for penetrating commentary on the manuscript. PES was supported by NJAES/USDA-32102 and NSF-



BSR-90-06589, TRM by NSF-BSR-83-14887 and NSF-BSR-90-06589.

## LITERATURE CITED

- ADAMS W. T., A. R. GRIFFIN and G. F. MORAN, 1992 Using paternity analysis to measure effective pollen dispersal in plant populations. *Am. Nat.* **140**: 762-780.
- AICKIN, M., 1984 Some fallacies in the computation of paternity probabilities. *Am. J. Hum. Genet.* **36**: 904-915.
- AVISE, J. C., R. A. LANSMAN and R. O. SHADE, 1979 The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *Peromyscus*. *Genetics* **92**: 279-295.
- BAIRD, M., I. BALAZS, A. GIUSTI, G. L. MIYAZAKI, L. NICHOLAS, *et al.*, 1986 Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. *Am. J. Hum. Genet.* **39**: 489-501.
- 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.
- BUDWOLE, B., R. CHAKRABORTY, A. M. GIUSTI, A. J. EISENBERG and A. J. ALLEN, 1991 Analysis of the VNTR locus D1S80 by the PCR followed by high resolution PAGE. *Am. J. Hum. Genet.* **48**: 137-144.
- BURKE, T., and M. W. BRUFORD, 1987 DNA fingerprinting in birds. *Nature* **327**: 149-152.
- CANNINGS, C., and E. A. THOMPSON, 1981 *Genealogical and Genetic Structure*. Cambridge University Press, New York.
- CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS, 1967 Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* **19**: 233-257.
- CHAKRABORTY, R., 1993 Generalized occupancy problem and its applications in population genetics, pp. 179-192 in *Genetic Variability in Human Diseases: Cells, Individuals, Families and Populations*, edited by C. F. SING and C. L. HANIS. Oxford University Press, New York.
- CHAKRABORTY, R., T. R. MEAGHER and P. E. SMOUSE, 1988 Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics* **118**: 527-536.
- CHARLESWORTH, D., S. W. SCHEMSKE and V. SORK, 1987 The evolution of plant reproductive characters; sexual *versus* natural selection, pp. 317-335 in *The Evolution of Sex and Its Consequences*, edited by S. C. STEARNS. Berkhauser Verlag, Basel.
- CLEGG, M. T., J. R. Y. RALSON and K. THOMAS, 1984 Chloroplast DNA variation in pearl millet and related species. *Genetics* **106**: 449-461.
- DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**: 1-38.
- DEVLIN, B., J. CLEGG, and N. C. ELLSTRAND, 1992 The effects of flower production on male reproductive success in wild radish populations. *Evolution* **46**: 10-30.
- DEVLIN, B., N. RISCH and K. ROEDER, 1992 Forensic inference from DNA fingerprints. *J. Am. Stat. Assoc.* **87**: 337-350.
- 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.
- ELLSTRAND, N. C., 1984 Multiple paternity within the fruits of the wild radish, *Raphanus sativus* L. *Am. Nat.* **123**: 819-828.
- ELLSTRAND, N. C., and D. L. MARSHALL, 1985 Interpopulation gene flow by pollen in wild radish, *Raphanus sativus*. *Am. Nat.* **126**: 606-616.
- ELSTON, R. C., 1986 Probability and paternity testing. *Am. J. Hum. Genet.* **39**: 112-122.
- ELSTON, R. C., and J. STEWART, 1971 A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**: 523-542.
- ESSEN-MÖLLER, E., 1939 Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. *Theoretische Brundlagen. Mitt. Anthrop. Ges. Wien* **68**: 9-53.
- FOLTZ, D. W., and J. L. HOOGLAND, 1981 Analysis of mating system in the black-tailed prairie dog (*Cynomys ludovicianus*) by likelihood of paternity. *J. Mammal.* **62**: 706-712.
- HAMRICK, J. L., and A. SCHNABEL, 1985 Understanding the genetic structure of plant populations: Some old problems and a new approach, pp. 50-70 In *Population Genetics in Forestry*, edited by H.-R. GREGORIUS. Springer-Verlag, New York.
- HANKEN, J., and P. W. SHERMAN, 1981 Multiple paternity in Belding's ground squirrel litters. *Science* **212**: 351-353.
- HEDRICK, P., 1992 Shooting the RAPDs. *Nature* **355**: 679-680.
- HILL, W. G., 1987 DNA fingerprints applied to animal and bird populations. *Nature* **327**: 98-99.
- HUGHES, C. R., and D. C. QUELLER, 1993 Detection of highly polymorphic microsatellite loci in a species with little allozyme polymorphism. *Mol. Ecol.* **2**: 131-138.
- JEFFREYS, A. J., and D. B. MORTON, 1987 DNA fingerprints of dogs and cats. *Anim. Genet.* **18**: 1-15.
- JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985a Hypervariable "minisatellite" regions in human DNA. *Nature* **314**: 67-73.
- JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985b Individual-specific "fingerprints" of human DNA. *Nature* **316**: 76-79.
- JEFFREYS, A. J., V. WILSON, S. L. THEIN, D. J. WEATHERALL and B. A. J. PONDER, 1986 DNA "fingerprints" and segregation analysis of multiple markers in human pedigrees. *Am. J. Hum. Genet.* **39**: 11-24.
- LEWIS, P. O., and A. A. SNOW, 1992 Deterministic paternity exclusion using RAPD markers. *Mol. Ecol.* **1**: 155-160.
- LI, C. C., and A. CHAKRAVARTI, 1985 Basic fallacies in the formulation of the paternity index. *Am. J. Hum. Genet.* **37**: 809-818.
- LYNCH, M., 1988 Estimation of relatedness by DNA fingerprinting. *Mol. Biol. Evol.* **5**: 584-599.
- LYONS, E. E., N. M. WASER, M. V. PRICE, J. ANTONOVICS, and A. F. MOTTEN, 1989 Sources of variation in plant reproductive success: a review of the theory. *Am. Nat.* **134**: 409-433.
- MCCRACKEN, G. F., and J. W. BRADBURY, 1977 Paternity and genetic heterogeneity in the polygynous bat *Phyllostomus hastatus*. *Science* **198**: 303-306.
- MEAGHER, T. R., 1982 The population biology of *Chamaelirium luteum*, a dioecious member of the lily family. IV. Two-sex population projections and stable population structure. *Ecology* **63**: 1701-1711.
- MEAGHER, T. R., 1986 Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely parents. *Am. Nat.* **128**: 199-215.
- MEAGHER, T., 1991 Analysis of paternity within a natural population of *Chamaelirium luteum*. II. Male reproductive success. *Am. Nat.* **137**: 738-752.
- MEAGHER, T. R., and J. ANTONOVICS, 1982 The population biology of *Chamaelirium luteum*, a dioecious member of the lily family. III. Life history studies. *Ecology* **63**: 1690-1700.
- MEAGHER, T. R., and E. A. THOMPSON, 1986 The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theor. Popul. Biol.* **29**: 87-106.
- MEAGHER, T. R., and E. A. THOMPSON, 1987 Analysis of parentage for naturally established seedlings of *Chamaelirium luteum*. *Ecology* **68**: 803-812.
- MICKEY, M. R., D. W. GJERTSON and P. I. TERASAKI, 1986 Empirical validation of the Essen-Moller probability of paternity. *Am. J. Hum. Genet.* **39**: 87-106.
- MILLIGAN, B. G., and C. K. McMURRY, 1993 Maximum likelihood

- analysis of male fertility using dominant and codominant genetic markers. *Mol. Ecol.* (in press).
- NAKAMURA, Y., M. LEPPERT, P. O'CONNELL, R. WOLFE, T. HOLM, *et al.*, 1987 Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**: 1616–1622.
- PALMER, J. D., R. A. JORGENSEN and W. F. THOMPSON, 1985 Chloroplast DNA variation and evolution in *Pisum*: patterns of change and phylogenetic analysis. *Genetics* **109**: 195–213.
- QUINN, T. W., and B. N. WHITE, 1987 Identification of restriction fragment length polymorphisms in genomic DNA of the lesser snow goose (*Anser caerulescens caerulescens*). *Mol. Evol. Biol.* **4**: 126–143.
- ROEDER, K., B. DEVLIN and B. G. LINDSAY, 1989 Applications 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.
- SCHOEN, D. J., and S. C. STEWART, 1987 Variation in male fertilities and pairwise mating probabilities in *Picea glauca*. *Genetics* **116**: 141–152.
- SMITH, D. B., and W. T. ADAMS, 1983 Measuring pollen contamination in clonal seed orchards with the aid of genetic markers, pp. 69–77 in *Proc. 17th Southern Forest Tree Improvement Conference*. University of Georgia, Athens.
- SMOUSE, P. E., and J. ADAMS, 1983 The use of linked genetic markers for paternity analysis, with special consideration of the HLA complex, pp. 518–522 in *Inclusion Probabilities in Parentage Testing*, edited by R. H. WALKER. Am. Assoc. Blood Banks, Arlington, Va.
- SMOUSE, P. E., and R. CHAKRABORTY, 1986 The use of restriction fragment length polymorphisms in paternity analysis. *Am. J. Hum. Genet.* **38**: 918–939.
- THOMPSON, E. A., 1975 The estimation of pairwise relationship. *Ann. Hum. Genet.* **39**: 173–188.
- THOMPSON, E. A., 1976a Inference of genealogical structure. *Soc. Sci. Inform.* **15**: 477–526.
- THOMPSON, E. A., 1976b Inference of genealogical structure. II. Quantifying genetic information. *Soc. Sci. Inform.* **15**: 491–506.
- THOMPSON, E. A., 1976c Inference of genealogical structure. III. The reconstruction of genealogies. *Soc. Sci. Inform.* **15**: 507–526.
- THOMPSON, E. A., 1986 *Pedigree Analysis in Human Genetics*. Johns Hopkins University Press, Baltimore.
- THOMPSON, E. A., 1987 Likelihood inference of paternity. *Am. J. Hum. Genet.* **39**: 285–287.
- THOMPSON, E. A., and T. R. MEAGHER, 1987 Parental and sib likelihoods and genealogy reconstruction. *Biometrics* **43**: 585–600.
- VALENTIN, J., 1984 Paternity index and attribution of paternity. *Hum. Hered.* **34**: 255–257.
- WETTON, J. H., R. E. CARTER, D. T. PARKIN and D. WALTERS, 1987 Demographic study of a wild house sparrow population by DNA fingerprinting. *Nature* **327**: 147–149.
- WILLIAMS J. G. K., A. R. KUBELIK, K. J. LIVAK, J. A. RAFALSKI and J. V. TINGEY, 1990 DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* **18**: 6531–6535.

Communicating editor: A. G. CLARK