Estimating Long-Term Mating Systems Using DNA Sequences

Brook G. Milligan

Department of Biology, New Mexico State University, Las Cruces, New Mexico 88003

Manuscript received March 4, 1993
Accepted for publication October 5, 1995

ABSTRACT

Plant mating systems often involve a mixture of self fertilizations and outcross fertilizations. The degree of selfing has a large impact on the genetic composition of natural populations and on the evolution of the mating system itself in response to such factors as inbreeding depression. This paper describes a means of estimating the long-term rate of self-fertilization from samples of alleles taken from individuals in a population. Use is made of the genealogy of pairs of alleles at a locus within individuals and pairs between individuals. The degree of selfing is closely related to the extent to which the number of nucleotide sites differing within an individual is reduced relative to the number differing between individuals. Importantly, the estimate of long-term selfing is largely independent of population size and is not affected by historical fluctuations in population size; instead it responds directly to the mating system itself. The approach outlined here is most appropriate to evolutionary problems in which the long-term nature of the mating system is of interest, such as to determine the relationship between prior inbreeding and inbreeding depression.

G E N E T I C transmission from one generation to the next is a major focal point in evolutionary biology because of its role in determining the genetic composition of a population. The nature of genetic transmission is mediated by the pattern of matings linking generations. In plants, the pattern of mating varies considerably and includes regular systems of inbreeding often with extensive self fertilization, breeding among relatives within small neighborhoods, and negative assortative mating maintained by incompatibility systems (CLEGG 1980). Plant populations rarely exhibit random mating (FRYNEILL 1957; WILLSON 1984). Characterization of mating systems is, therefore, of prime concern in the study of plant population genetics and evolution.

Quantification of plant mating systems often involves a mixed mating model in which matings consist of either selfing or outcrossing to a common pollen pool (CLEGG 1980; RITLAND 1984; BROWN 1989). The degree of selfing estimated by this model positions a population along a continuum between complete selfing and random mating. That continuum is central to much of the theory concerned with the evolution of mating systems (LLOYD 1979; LANDE and SCHEMSKE 1985). As such the mixed mating model provides a useful summary of many mating systems, although more complex mating models (RITLAND 1984; SCHOEN and CLEGG 1984; ROEDER et al. 1989) may be appropriate at times.

Estimation of the degree of selfing has traditionally been based on surveys of allele segregation in sets of progeny arrays collected from a population (BROWN and ALLARD 1970; BROWN et al. 1975, 1986, 1989; WARD 1989; HOLTSFORD and ELLSTRAND 1990). This approach offers the advantages of being experimentally tractable and efficient at quantifying mating systems for a single generation. However, because of the numerous demographic and ecological factors known to affect the mating process (LLOYD 1980; WILLSON 1984; BROWN 1989; BROWN et al. 1989), single generation estimates may not be useful for studies concerned with the long-term evolution of mating systems. For example, single generation estimates of selfing rate made over multiple years demonstrate temporal variation (VASER and HARDING 1976; HAMRICK 1982; HOLTSFORD and ELLSTRAND 1990; DOE and RITLAND 1993). Given short-term temporal variation in breeding system, a measure that provides an indication of the long-term mating system may be useful in the context of evolutionary studies because the time scales of the phenomena of interest are more closely matched. In this paper I show that such a measure may be calculated from samples of DNA sequences taken from a population. In particular, I rely on the genealogical information contained within DNA sequences to infer both the selfing rate and the mutation rate. In so doing I show that DNA sequences are especially useful in the study of mating systems. Further, the necessary data may be readily obtained using modern PCR-based techniques for screening natural variation in DNA sequences (LESSA 1992; LESSA and APPLEBAUM 1993; LEEBENS-MACK 1995; STRAND et al. 1995).

PARTIAL SELFING IN A FINITE POPULATION

The traditional mixed mating or partial selfing model of mating randomly classifies gamete unions into two classes: those derived from selfing and those derived...
from outcrossing to a uniform pollen pool. This reflects the fact that individuals typically exhibit mixed mating strategies; however, under all conditions for which the coalescent approximations are valid the model is also formally equivalent to one in which fixed selfing or outcrossing mating strategies are assigned randomly to individuals. The primary parameters that describe the traditional partial selfing model are the selfing rate $s$ and the pollen pool allele frequency $p_i$ for the $i$th allele. These parameters may be estimated from information on genotype frequencies in collections of progeny (Brown and Allard 1970; Cheljac et al. 1983; Ritland 1986). Although the maternal genotypes are not required, estimation is simpler if they are known (Brown et al. 1975; Brown 1989).

In contrast to genotype frequencies, DNA sequences contain information on the genealogical relationships of the sampled alleles. This is because the number of nucleotide sites differing between a pair of sampled alleles depends on the time at which they were derived from a common ancestral allele in the past (Hudson 1990), that is, the time at which their ancestry coalesces. The coalescence times for a sample of alleles are dependent on demographic factors in the population. As a result, the number of nucleotide sites differing between alleles may be used to estimate demographic factors such as effective population size or migration rate (Slatkin 1991; Felsenstein 1992), and genetic parameters such as the mutation rate or the recombination rate (Hudson 1990; Felsenstein 1992). The same genealogical framework is applicable to the problem of estimating mating systems from DNA sequences.

Throughout I will consider a sample of DNA sequences for distinct alleles at a single diploid locus, collected in such a way that the identity of the individual containing each allele is known. The number of nucleotide sites differing between pairs of alleles in the sample provides the genealogical information we seek. It is important, however, to distinguish pairs of sampled alleles corresponding to the two within a single individual from other pairs in which each allele is derived from a distinct individual. This distinction allows estimation of the mating system as well as the mutation rate. Clearly, the approach outlined here can be extended to incorporate information on the entire genealogy of alleles contained in the sample; that, however, is beyond the scope of this paper.

To develop a specific model to use in estimating the mating system, I follow the classical mixed mating model introduced above (Clegg 1980; Brown 1989) with modifications to account for the nature of DNA sequence evolution in finite populations. In this context, the central parameters of the model are the probabilities of identity by descent for pairs of alleles. Because pairs sampled within versus between individuals are distinguished, two identities must also be distinguished. Let $f_{12}(t)$ be the probability that the two alleles within a single individual are identical by descent in generation $t$. Likewise, let $f_{13}(t)$ be the probability that two alleles sampled from distinct individuals are identical by descent in generation $t$. Finally, let $N$ be the census population size. Although $N$ is assumed to be constant through time, this may be relaxed to the assumption that the population is never very small because the main results are largely independent of $N$.

Given union of gametes through selfing, the probability of identity by descent within an individual increases each generation (Crow and Kimura 1970, equation 3.8.1):

$$f_{12}(t + 1) = \frac{1}{2} + \frac{1}{2} f_{12}(t) . \quad (1)$$

Likewise, the probability of identity by descent between individuals increases under selfing each generation (Crow and Kimura 1970, equation 3.11.3):

$$f_{13}(t + 1) = \frac{1}{2N} + \frac{1}{2} f_{13}(t) + \left( 1 - \frac{1}{N} \right) f_{22}(t) . \quad (2)$$

While these probabilities of identity by descent may provide the natural framework for deriving the genetic effects of consanguineous matings, it is easier to develop the genealogical model in terms of the probabilities of nonidentity. Therefore, consider $g_{12}(t) = 1 - f_{12}(t)$ and $g_{13}(t) = 1 - f_{13}(t)$, the corresponding probabilities of nonidentity by descent. These terms represent the probability that two alleles do not have a common ancestor by generation $t$ in the past. In general, recursions for the probabilities of nonidentity are easily derived from the recursions for the corresponding probabilities of identity (Slatkin 1991). In the mixed mating model under consideration, recursions for $g(t)$ are

$$g_{12}(t + 1) = \frac{1}{2} g_{12}(t) \quad \text{and} \quad (3)$$

$$g_{13}(t + 1) = \frac{1}{2N} g_{13}(t) + \left( 1 - \frac{1}{N} \right) g_{22}(t) . \quad (4)$$

Note that, as with all recursions in coalescent problems, the labeling of generations by time differs somewhat from that encountered with traditional prospective population genetics even though in both cases the recursions progress forward in time. In (3) and (4) $t$ refers to the number of generations in the past during which the alleles have been distinct.

Next consider the union of gametes through outcrossing to a pollen pool derived from a finite population of constant size $N$. This follows the standard Wright-Fisher model of random mating in a finite population (Fisher 1930; Wright 1931). In this case the increase in both components of identity by descent is given by the right-hand side of (2), and the decrease in both components of nonidentity by descent is given by the right-hand side of (4).

To simplify the following algebra, define a vector of
the probability of nonidentity as $g(t) = [g_u(t), g_o(t)]^T$. The changes in this vector during a single generation of selfing and during a single generation of outcrossing may be described by the matrices $S$ and $T$, respectively:

$$S = \begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{2N} & 1 - \frac{1}{N} \end{pmatrix}$$

and

$$T = \begin{pmatrix} \frac{1}{2N} & 1 - \frac{1}{N} \\ \frac{1}{2N} & 1 - \frac{1}{N} \end{pmatrix}.$$  

(5)

The overall change in nonidentity due to a mixture of selfing with probability $s$ and outcrossing with probability $1-s$ is given in matrix form as

$$g(t+1) = (sS + (1-s)T)g(t).$$

(7)

Finally, for convenience let $A$ represent the transition matrix $sS + (1-s)T$.

The recursion given by (7) fully identifies the changes from one generation to the next in nonidentity that result from partial selfing. It is assumed that initially all alleles were unrelated, so $g(0) = (1, 1)^T$. Because $g(t)$ gives the probability of noncoalescence by generation $t$, the probability that coalescence occurs at generation $t$ in the past is given by (SLATKIN 1991)

$$p(t) = g(t) - g(t+1).$$

(8)

From these recursions, one can calculate the complete distribution of coalescence times for pairs of alleles taken from either within or between individuals.

MEAN COALESCENCE TIMES AND THE NUMBER OF SEGREGATING SITES

From samples of pairs of alleles taken either within individuals or between individuals one may estimate for a population the number of nucleotide sites by which the pairs differ, that is, the number of segregating sites (WATTERSON 1975). Under the infinite sites model of DNA sequence evolution, the expected number of segregating sites is simply related to the average time until ancestry for a pair of alleles coalesces (HUDSON 1990). Let $\bar{t} = (\bar{t}_u, \bar{t}_o)^T$ be the average time of coalescence for alleles sampled within and between individuals. In terms of the above recursion for $g(t)$, the mean coalescence time is given by (SLATKIN 1991)

$$\bar{t} = \sum_{i=1}^{\infty} tP(t) = (I - A)^{-1}g(0),$$

(9)

where $I$ is an identity matrix. For the mixed mating model under consideration, the mean coalescence time is

$$\bar{t} = 2N \begin{pmatrix} 1 - \left(1 - \frac{1}{N}\right)s \\ 1 - \frac{1}{2} \left(1 - \frac{1}{N}\right)s \end{pmatrix}.$$  

(10)

Note that the ratio of coalescence times for alleles within vs. between individuals is less than one for $s > 0$. As one might expect, the presence of selfing leads to a shorter time to coalescence for alleles within an individual than for alleles sampled between individuals.

Mutations accumulate along the entire lineage connecting each pair of sampled alleles. Assume that all mutations are new to the population and that they occur at the rate $\mu$ per gene per generation. Because the expected length of the lineage between two alleles is $2\bar{t}$, the expected number of nucleotide sites differing between pairs of alleles is given by $2\bar{t}\mu$, and the average number of sites segregating in pairs of alleles is

$$\bar{y} = \theta \begin{pmatrix} 1 - \left(1 - \frac{1}{N}\right)s \\ 1 - \frac{1}{2} \left(1 - \frac{1}{N}\right)s \end{pmatrix},$$

(11)

where $\theta = 4N\mu$ is the mutation parameter of neutral models of evolution (KIMURA 1983). After some rearrangement of terms one obtains the following method of moments estimators of the selfing rate $s$ and mutation parameter $\theta$:

$$\hat{s} = 2 - \frac{N}{N-1} \frac{\bar{y}_s - \bar{y}_o}{2\bar{y}_s - \bar{y}_o}$$

and

$$\hat{\theta} = 2\bar{y}_s - \bar{y}_o.$$  

(12)

(13)

Observations of the number of nucleotide sites differing between pairs of alleles sampled within ($\bar{y}_w$) and between ($\bar{y}_b$) individuals may be used to estimate both the degree of selfing and the central parameter of neutral models, $\theta$. Note that because of the selfing component, neither traditional estimators of $\theta$ (EWENS 1972; WATTERSON 1975; NEI and TAJIMA 1981; TAVARE 1984) nor modern coalescent derivatives (STROBECK 1983; FELSENSTEIN 1992; FU 1994a,b; KUHNER et al. 1995) can be used in the context of mixed mating; instead, the mutation parameter $\theta$ must be jointly estimated with the selfing rate $s$. As a point of comparison, the coalescent estimators derived above will be contrasted with a traditional estimator of $\theta$ based on the distribution of segregating sites (WATTERSON 1975)

$$\hat{\theta}(F) = \frac{1 - F}{F},$$

(14)
where \( F \) is the frequency of individuals homozygous for alleles identical by descent.

One alternative long-term measure of selfing rate is based on the relationship between the equilibrium value of homozygosity, \( F \), and selfing rate (Brown and Allard 1970),

\[
\delta(F) = \frac{2F}{1 + F^2}. \tag{15}
\]

This estimator shares several properties with the coalescent one presented here and is therefore of interest as a point of comparison. The most important property in common is that both are equilibrium measures and are therefore indicative of the long-term breeding system. Both are in contrast to short-term measures based on the segregation of alleles during a single generation of mating.

**Properties of the Distribution of Segregating Sites**

All of the information needed to describe the behavior of these estimators of the selfing rate and the mutation parameter is encapsulated by the distributions of segregating sites within and between individuals. Each of these distributions is simply a weighted mixture of many Poisson distributions of mutation, one for each generation \( t \) and with mean \( 2\mu t \). The weights are given by the probability of coalescence at time \( t \) found from (8) and a series of iterations of (7) (Hudson 1990). Generally, these distributions will depend on the selfing rate \( s \) and the mutation parameter \( \theta \). Figure 1 illustrates these distributions for \( \theta = 5 \), but for any given value of \( \theta \), the distributions under mixed mating are intermediate between those for continual selfing and those for random mating in a finite population.

While the entire distribution of segregating sites is useful and may provide the basis for a test of the underlying partial selfing model, of primary interest with regard to the mating system estimators are the moments of the distribution and the degree of homozygosity (Figure 2). Both the mean and variance of the distributions decline with increasing selfing, with the greatest decline being in the moments of the within-individual distribution. In contrast, the degree of homozygosity increases with increasing selfing, as indeed it should. Note that with complete outcrossing (\( s = 0 \)) there is still a substantial probability of identity by descent within individuals due to the finite population size, and that there is a strong dependence of that probability on population size.

The means of the two distributions of segregating sites are sufficient to calculate coalescent estimates of both the selfing rate and the mutation parameter [(12) and (13)], whereas the probability of no segregating sites within individuals \( (F) \) is sufficient to calculate both traditional estimates [(15) and (14)]. As a result, the statistical properties of the estimators can be determined from a set of complete distributions for the number of segregating sites, each generated as for Figure 1 under a variety of conditions representing different selfing rates, mutation rates, and population sizes.

**Bias of the estimates:** The degree of bias of the estimators is evident from a comparison between the estimates obtained and the parametric values used to generate the complete distributions of segregating sites (Figure 3). It is evident that the coalescent estimator of selfing rate developed here exhibits essentially no bias over a wide range of population sizes throughout the range of selfing rates. In contrast, the only alternative estimator of long-term mating system, one based on homozygosity, exhibits extreme bias especially under conditions of predominant outcrossing. This is because finite populations develop a significant degree of homozygosity even in the absence of selfing. Thus, this bias is likely to be less of a problem for larger populations but may be exacerbated by fluctuations in population size. The coalescent estimator is not influenced by population size, even by fluctuating population size, and is therefore more useful over a wider range of conditions in natural populations.

Figure 3 also illustrates that the coalescent estimator...
Estimating Mating Systems

10.0

T:;;:.,.,.

12,500

---

N = 2,500

8.0

-------

Between 25,000...

Within....

6.0

-------

Between...

Within....

4.0

Within and Between

2.0

0.0

0.0 0.2 0.4 0.6 0.8 1.0

Selfing Rate

120

I

I

1 I I

I

I

I

I

1

80

-------

Between...

Within....

60

-------

Within and Between

40

-------

Within and Between

20

-------

Within and Between

0.0

0.0 0.2 0.4 0.6 0.8 1.0

Selfing Rate

FIGURE 2.—Mean (top) and variance (middle) of the distributions of segregating sites. The lower line of each pair gives the moments for samples within an individual; the upper line gives the moments for samples between individuals. The fraction of individuals with alleles identical by descent (bottom), i.e., lacking mutations entirely, is also given. $\mu = 10^{-4}$.

of the mutation parameter is unbiased over the full range of selfing rates. In distinct contrast, the traditional estimator of $\theta$ that does not account for mixed mating (14) exhibits rapidly increasing bias with increasing selfing rate, even though it exhibits no bias when mating is random. The fact that traditional estimators of the mating system and the mutation parameter are severely biased under a wide range of conditions greatly restricts their utility for the study of mixed mating populations, such as those characteristic of most plants.

Although Figure 3 illustrates that neither coalescent

estimator is asymptotically biased, their behavior under conditions of finite sample size may differ. Based on a series of simulations in which the mean estimate from 100,000 samples of various sizes, ranging from 20 to 320 pairs of alleles within and between individuals, taken from the same set of distributions used above, it is evident that the coalescent estimator of the mutation parameter is essentially unbiased over a broad range of sampling conditions. The estimator of selfing rate, however, exhibits some downward bias that increases with both decreasing selfing rate and decreasing sample size but decreases with increasing mutation rate. For example, the downward bias in selfing rate is 7% when $s = 0.5$, $\theta = 1$, and sample size is 160, but increases to 30 and 40% when selfing rate is reduced to $s = 0.2$ or sample size is reduced to 20, respectively, and decreases to 2.6% when $\theta = 10$.

Variance of the estimates: The estimates of selfing rate $s$ and mutation parameter $\theta$ may vary as a result of the finite nature of the sample of alleles available. An estimate of that variation may be achieved from the

FIGURE 3.—Coalescent and traditional estimates as a function of the parametric values of selfing rate and population size $N$. Equality between the parametric and estimated values of selfing rate (top) is represented by the solid diagonal line coinciding with the coalescent estimator; equality for values of the mutation parameter (bottom) is represented by the series of horizontal lines coinciding with the coalescent estimator. Note that the line depicting the coalescent estimator of selfing rate represents superimposition of values for the same range of population sizes as for $\delta(F)$. In all cases, $\mu = 10^{-4}$. 

Variance of the estimates: The estimates of selfing rate $s$ and mutation parameter $\theta$ may vary as a result of the finite nature of the sample of alleles available. An estimate of that variation may be achieved from the
same finite samples used to investigate bias. Figure 4 illustrates the variation in the estimates of selfing rate $s$ and mutation parameter $\theta$ as functions of both the selfing rate and $\theta$. The variance in $s$ depends much more strongly on selfing rate than the variance in $\theta$, though both decline with increasing selfing rate.

The variance of both estimators decreases dramatically as $\theta$ increases. This point bears on the choice of markers to use when estimating mating system. Markers that exhibit higher rates of mutation will provide less variable estimates of both selfing rate and $\theta$; as noted above they will also provide less biased estimates.

In addition to choice of markers, one has control over the number of samples to include in a study. Figure 5 illustrates how the variability depends on sample size. All measures of variation decline with increasing selfing or increasing sample size. Clearly rather large samples, on the order of 100 or more pairs of alleles within and between individuals, are necessary to obtain reasonably accurate estimates. This may appear to make impossible the practical application of these estimators. However, samples of that size are easily within the range of modern PCR-based techniques that are able to distinguish alleles at a single locus (LESSA 1992; LESSA and APPLEBAUM 1993; LEEBENS-MACK 1995; STRAND et al. 1995).

It is also important to understand that the estimates of variation illustrated in Figures 4 and 5 represent upper bounds; the actual degree of variation in samples taken from natural populations is likely to be much lower. The simulations used to generate the figures iterated (7) to give the distributions from which finite samples were drawn. This makes the assumption that each pair of alleles drawn from a distribution is independent of the remainder and that all alleles within individuals are independent of all alleles between individuals. This is not strictly true, because the alleles are related through the larger genealogy of the entire population. The effect of that larger genealogy is to reduce the actual variation between distinct pairs of alleles and therefore to reduce the variation in the estimates derived from those pairs. The effect can be substantial, so in practice much smaller samples than suggested by Figure 5 or much less variable markers than suggested by Figure 4 can be used.

**Approach to equilibrium:** The estimates of mating system are based on the assumption of an equilibrium with regard to degree of inbreeding in the population. Since natural populations may not be at an equilibrium, it is important to determine how rapidly the expected coalescence time $\tau$ approaches its equilibrium. The rate
of approach of coalescence time determines the rate of approach of the number of segregating sites and hence the mating system estimate.

The dominant eigenvalue of $A$ determines the rate of approach of $I$ to its equilibrium value. This eigenvalue is given by

$$\lambda = s \lambda_s + (1 - s) \lambda_u + \delta,$$

(16)

where $\lambda_s = 1 - 1/N$ and $\lambda_u = 1 - 1/(2N)$ are the dominant eigenvalues under continual selfing and under outcrossing described by the Wright-Fisher model of random mating in a finite population, respectively. The final term, $\delta$, is a small correction factor ($0 \leq \delta < 1/(4N)$) given by

$$\delta = \frac{1}{2} \sqrt{\frac{1}{4N^2} (2N - 1 + Ns - s)^2 - 2 \frac{N - 1}{N} s - \left( \frac{1}{4} - \frac{1}{2N} \right) s - \left( \frac{1}{2} - \frac{1}{4N} \right) (1 - s) \cdot (17)$$

For reasonable population sizes, the dominant eigenvalue under partial selfing is intermediate between that under continual selfing and that under random mating.

Under complete selfing the time to coalescence is on the order of $N$ generations; with random mating it is on the order of $2N$ generations. Partial selfing leads to intermediate coalescent times. As a result, the coalescent estimators of both the selfing rate and the mutation parameter will respond to transient conditions occurring over relatively long periods of time in the past, unlike traditional estimators dependent only on homozygosity that equilibrates much more rapidly. This feature is advantageous for evolutionary biologists because the estimates obtained are representative of the evolutionary time frame of interest.

**DISCUSSION**

The widespread awareness of the importance of plant mating systems in determining the genetic composition of natural populations and the recognition of a diversity of mating systems exhibited by different species has focused the interest of evolutionary biologists and geneticists on the problem of quantifying mating systems. The traditional approach has relied on the mixed mating model of partial selfing together with genetic assays of variation at allozyme loci. While that traditional approach has yielded valuable insight into plant mating systems, it provides only a short-term view. The traditional approach of the number of segregating sites and hence the mating system estimate.

The estimators described here demonstrate that DNA sequences may be used to quantify the parameters of the mixed mating model in a finite population. The primary observations needed are the mean number of nucleotides differing between the two alleles at a diploid locus sampled from within individuals and the mean number of nucleotides differing between pairs of alleles sampled from distinct individuals. Both quantities may be estimated from a sample of alleles as long as the identity of the individuals sampled is known and both alleles are sampled from at least some individuals. At first, this sampling scheme may appear daunting due to the need to separately identify both alleles at a single locus. Several gel-based screening procedures now allow such separation however (Lessa and Applebaum 1993), so the necessary experimental techniques have already been developed.

Given samples of both alleles from within a number of individuals and pairs of alleles from distinct individuals, the estimated fraction of matings due to selfing and the mutation parameter, $\theta = 4N\mu$, governing the neutral theory of evolution (Kimura 1983) can be estimated. Both estimates depend on the degree to which the number of segregating sites is reduced within an individual compared with between individuals. A greater reduction in the number of sites differing within an individual results from a higher degree of selfing in the population and from a larger mutation rate.

The estimate of long-term selfing rate is essentially unbiased over the full range of actual selfing rates for a wide variety of mutation rates and population sizes. This is in distinct contrast to the only other measure of long-term selfing rate, that based on the equilibrium value of homozygosity. Because identity by descent depends strongly on population size, even in the absence of selfing, this estimator exhibits a large and highly variable bias. Likewise, the estimate of the mutation parameter $\theta = 4N\mu$ is unbiased over the same range of parameters. This is also in distinct contrast to traditional measures of $\theta$ that assume random mating and exhibit severe bias under other mating systems.

Although unbiased, the genealogical estimators presented here exhibit relatively large upper bounds on the variance unless samples include pairs from at least 100 individuals. Taking advantage of gel-based screening procedures such as those based on differences in physical properties due to differences in sequence (Lessa and Applebaum 1993), one can readily amass appropriate data sets of this size. Thus, while the variance may at first appear impossibly large, in fact, the techniques are currently available and the time required for genotyping a sample of 100 individuals is
relatively short, on the order of 2 wk. Further, when a choice of markers is available, the variance may be reduced by relying on those with higher mutation rates; in some cases this may obviate the need for large samples. Reasonable estimates are within the realm of current technology. Importantly, the estimates of variation presented are upper bounds and may be significantly greater than the actual variation. As a result, much smaller samples or less variable markers than suggested above are likely to be useful in practice.

The reduction of variance with increased marker mutation rate is reflective of a general property of coalescent estimators of demographic properties of populations. Mutations serve simply to mark different portions of the genealogy relating sampled alleles to each other but occur entirely independently of the demographic events that influence the temporal structure of the genealogy. In the case of the selfing rate estimates discussed here, mutations occur independently of the mating events that shape the times to common ancestry of pairs of sampled alleles. However, higher mutation rates provide more events to mark different parts of the genealogy and therefore provide greater resolution of the past history and a reduced variance. In fact, markers with higher mutation rates will generally lead to estimates of lower variance, because these features are true of most coalescent estimators of demographic properties.

The most useful estimators of selfing rate are those that are not sensitive to historical fluctuations in population size but instead respond directly to the mating system itself. Stochastic variation in population size had little effect on the mean coalescence times observed in this model and hence on the estimate of selfing rate, based on simulations in which the complete distribution of coalescence times was calculated. This point is reinforced by the observations that only a 1% error will be introduced by neglecting population size in (12) if \( N = 100 \) and that the points illustrated in Figure 3 are entirely superimposed, even though population size varies over an order of magnitude. Thus, if the population size is never small, the assumption of constant population size used in deriving the estimators is of little import.

Finally, it should be emphasized that the mating system estimates obtained from this model are long-term averages, not characterizations of the mating events in the previous generation. Further, they are not the arithmetic mean of past selfing rates, for the same reason that effective population size is not the arithmetic mean of past population sizes. In contrast, estimates obtained from the classical mixed mating model, the effective selfing model, or other models that examine the segregation of markers from one generation to the next (Ritland 1984, 1986; Milligan and McMurry 1993) quantify a single generation of mating but do not include the effects of long-term fluctuation in the mating system. These two approaches are therefore useful for different types of investigation. The latter is most appropriate for identifying specific mating events, perhaps with the idea of relating them to pollinator behavior or response to plant traits. In this case, the ability of the mating system is important to capture. The former is most appropriate for evolutionary studies of the long-term mating system when the vagaries of year to year variation might interfere with accurate predictions. For example, studies of the relationship between inbreeding depression and mating system (Holtsford and Ellstrand 1990; Latka and Ritland 1994) might benefit from a long-term view of the mating system rather than one that may differ from year to year.

Additional means of categorizing DNA sequence variation within and among individuals in natural populations will provide the basis for more complex models of plant mating systems. Use of the joint distribution of variation among more than two sequences will provide additional information and will undoubtedly lead to mating system estimators with less variance. The results obtained here, however, demonstrate that genealogical analysis of sequence variation is a fruitful approach to evolutionary and genetic problems in plant population biology.

Thanks are due to Yun-Xin Fu, Michael Lynch, James Leebens-Mack, Rick Miller, Kermit Ritland, Allan Strand, and several anonymous reviewers for helpful discussions or comments on earlier drafts. This project was supported in part by National Science Foundation grant BSR 89-06129.

LITERATURE CITED


Estimating Mating Systems


LEEBENS-MACK, J., 1995 An Investigation of the Consequence of Natural Hybridisation Between Two East Texas Baptisia (Fabaceae) Species, Ph.D. Thesis, University of Texas, Austin.


Communicating editor: M. LYNCH