Extensions of models for the estimation of mating systems using n independent loci

K Ritland

Department of Forest Sciences, University of British Columbia, Vancouver, BC V6T1Z4, Canada

Inferences about plant mating systems increasingly use highly informative genetic markers, and investigate finer facets of the mating system. Here, four extensions of models for the estimation mating systems are described. (1) Multiallelic probabilities for the mixed selfing-random mating model are given; these are especially suitable for microsatellites; a generalized Kronecker operator is basis of this formula. (2) Multilocus probabilities for the 'correlated-matings model' are given; interestingly, comparisons between single- vs multilocus estimates of correlated-paternity can provide a new measure of population substructure. (3) A measure of

biparental inbreeding, the 'correlation of selfing among loci', is shown to approximate the fraction of selfing due to uniparental (as opposed to biparental) inbreeding; also joint estimation of 1- 2- and 3-locus selfing rates allow separation, under a simple model, of the frequency vs the magnitude of biparental inbreeding. (4) Method-of-moments estimators for individual outcrossing rates are given. Formulae are given for both gymnosperms and angiosperms, and the computer program 'MLTR' implements these methods.

Heredity (2002) **88**, 221–228. DOI: 10.1038/sj/hdy/6800029

Keywords: outcrossing; inbreeding; mating system; relatedness; genetic markers

Introduction

The marker-based estimation of plant mating systems dates from the seminal studies of Jones (1916) and Fyfe and Bailey (1951). Their studies relied upon the pattern of genetic transmission from parents to progeny as shown by genetic markers. This pattern can be used to infer retrospectively the mating system (Clegg, 1980). Retrospective studies of plant mating systems have become a mainstay of plant population studies (Barrett and Eckert, 1990).

Plant mating systems are often described by the 'mixed mating model', where a fraction of progeny are derived from self-fertilization and the remainder derived from outcrossing at random. Many improvements and elaborations of the original mixed mating model have been made. Brown and Allard (1970) have shown how maternal parentage could be inferred from progeny arrays, whilst Shaw *et al* (1980) and Ritland and Jain (1981) introduced multilocus procedures for estimation of outcrossing. Compared with single-locus methods, multilocus methods have lower statistical variance and higher robustness against the violation of model assumptions (Ritland and Jain, 1981), whilst comparisons of single with multilocus estimates indicate levels of mating between relatives (Shaw *et al*, 1980).

With advances in both marker methods and statistical inference, more detailed facets of mating systems, such as patterns of paternity and dispersal, have been modeled and inferred (Cruzan, 1998). For example, while usually

progeny from common mothers are assumed to be constant mixtures of half-sibs and selfed progeny, it may be assumed progenies are full-sibs (Schoen and Clegg, 1984). And, whilst most studies have estimated selfing rates as population averages, we should remember that mating systems are dynamic quantities subject to environmental effects and may differ among individuals within populations. To this end, estimation procedures for outcrossing rates of individual plants have been developed (Cruzan and Arnold, 1994; Ivey and Wyatt, 1999).

Here, I provide four extensions of current estimation procedures along these lines. (1) Multiallelic probabilities for the probability of offspring genotype, given parent genotype, are defined for the mixed mating model. They are based on the Kronecker operator, which was used to describe estimators for coefficients of relatedness (Ritland, 2000). Expressions are first developed for the case where the gametophyte is assayed (as in conifers), then extended to the angiosperm case (or to animals, such as partially selfing snails). These probabilities accommodate the large numbers of alleles typical of highly informative microsatellite loci. (2) Multilocus formulae are then given for the correlated-matings model (Ritland, 1989), a model which characterizes shared paternity among family members. Multilocus estimation of correlated matings gives estimates of lower statistical variance than the single-locus estimates as used by Ritland (1989). Interestingly, the difference between single-locus *vs* multilocus estimates of correlated-matings can provide a new measure of population substructure. (3) Two ways of characterizing biparental inbreeding are introduced, both based upon the phenomena that under biparental inbreeding, the multilocus selfing rate depends upon the number of loci, decreasing with greater number of loci. These two ways involve: (3a) differences between one-

Correspondence: K Ritland, Department of Forest Sciences, University of British Columbia, Vancouver, BC V6T1Z4, Canada.

E-mail: kermi.ritland@ubc.ca

Received 14 May 2001; accepted 29 October 2001

locus, two-locus, and three-locus selfing rates and (3b) the correlation of selfing of locus-pairs within individuals. The correlation of selfing among loci approximates the fraction of selfing caused by uniparental (as opposed to biparental) inbreeding. (4) 'Method-of-moments estimators' (MMEs) of individual outcrossing rates are derived. These are analogous to MMEs for relatedness as derived in Ritland (1996), and are of use in fine-scale studies of mating systems, as they are unbiased at small sample sizes when pollen gene frequency is not jointly estimated. Statistical properties of each extension are discussed. These methods are implemented in a new estimation program, 'MLTR'.

Extensions of mating system models

The probability model underlying multilocus estimation of mating system assumes n unlinked loci. In the mixed-mating model, progeny are either selfed or randomly outcrossed. Outcrossing also occurs at random to a pollen pool at linkage equilibrium. The effect of these assumptions has been examined in the case of pollen pool heterogeneity (Smyth and Hamrick, 1984) and linkage (Hedrick and Ritland, 1990).

Estimates of selfing rate are found by maximizing the likelihood of the data with respect to selfing rate (and pollen gene frequencies, if also estimated), using numerical methods such as Newton-Raphson or the expectation-maximization (EM) method (Cheliak *et al.*, 1983). Maternal parentage can be inferred by computing likelihoods of entire progeny arrays across possible maternal genotypes, and choosing the parent genotype giving the highest likelihood; however this requires assuming either that families are full sibs (Schoen and Clegg, 1984) or half-sibs (Brown and Allard, 1970). Errors of estimates can be found with the bootstrap method, where entire progeny arrays are resampled (Ritland, 1990).

Multiallelic probabilities based upon the Kronecker operator

Previously, probabilities of offspring under the multilocus mixed mating model of Ritland and Jain (1981), as well as the correlated-mating model of Ritland (1989), were given for only diallelic loci, and computer programs (Ritland, 1990) were written for at most three alleles. This is a hindrance when markers are highly polymorphic, such as microsatellites. Furthermore, the 'conifer' model (when the haploid maternal contribution to the zygote can be assayed) has been separately treated from the usual 'angiosperm' model (Ritland and El-Kassaby, 1985). Yet, the latter model can be derived as a mixture model of the former, and illustrates how models for incomplete data are formed.

Multiallelic probabilities are best represented by the 'Kronecker operator', δ . This operator allowed compact expression for estimators of pairwise relatedness (Ritland, 2000). For mixed-mating probabilities, first consider a single-locus. The Kronecker operator is defined such that if two alleles A_i and A_j are the same (eg, the same band or sequence), then $\delta_{ij} = 1$, while if different, $\delta_{ij} = 0$. Now, for any progeny allele k , define

$$D_k^{ij} = \left(\frac{\delta_{ik} + \delta_{jk}}{2} \right) \quad (1)$$

which is the probability that allele k is transmitted to the

progeny, given parent $A_i A_j$. This expression will be the backbone of the formulae that follow.

Gametophyte-known (conifer): In the gametophyte-known case, progeny alleles can be assigned to eggs *vs* pollen. Let allele k in the progeny be the megagametophyte (derived from the mother) so that allele l is derived via pollen (note that $k > l$ is possible). Let P_{kl}^{ijs} be the probability of observing progeny genotype $A_k A_l$ given that the parent is $A_i A_j$ and self-fertilizes, and let P_{kl}^{ijt} be the probability of observing progeny $A_k A_l$ given the parent is $A_i A_j$ and outcrosses (s denotes selfing, t denotes outcrossing). Then,

$$P_{kl}^{ijs} = D_k^{ij} D_l^{ij} \quad (2)$$

$$P_{kl}^{ijt} = D_k^{ij} p_l$$

where the frequency of alleles in the population is indicated by subscripted p . P_{kl}^{ijs} assumes regular segregation in both sexes and random fusion of gametes. P_{kl}^{ijt} assumes that the probability of encountering allele l in the pollen is simply the frequency of l in the population, p_l ; thus outcrossing is assumed to occur to a homogenous pollen pool.

Gametophyte-unknown (angiosperm): In this case, progeny alleles cannot be assigned to parents by sex. Reflecting the 'incomplete' nature of the data, the probability of an outcrossed progeny genotype is the arithmetic average of the probabilities of the two alternative paternal alleles, k and l . As now $k \leq l$ always, the heterozygote probability is multiplied by 2, and any probability is multiplied by $(2 - \delta_{kl})$. The probabilities of progeny, conditioned on selfing versus outcrossing, are thus

$$P_{kl}^{ijs} = (2 - \delta_{kl}) D_k^{ij} D_l^{ij} \quad (3)$$

$$P_{kl}^{ijt} = \frac{1}{2} (2 - \delta_{kl}) (D_k^{ij} p_l + D_l^{ij} p_k)$$

Multilocus probability of offspring: If s is the rate of selfing, and $t = 1 - s$ is the rate of outcrossing, the multilocus likelihood of an offspring is:

$$s \prod_{\text{loci}} P_{kl}^{ijs} + t \prod_{\text{loci}} P_{kl}^{ijt} \quad (4)$$

(subscripts denoting loci are omitted for brevity). These probabilities are then incorporated into procedures for estimating selfing via progeny arrays. While these and the following formulae assume that maternal parentage is known, these formulae are readily incorporated into procedures that infer maternal parentage from the progeny array genotypes (eg, Brown and Allard, 1970). Statistical tests that can be used for such indirect estimation of mating system parameters include construction of confidence intervals via inversion of the information matrix (Ritland and Jain, 1981) or bootstrapping (Ritland, 1990), or performing a likelihood ratio test (comparing the likelihood of estimated values with the likelihood under their null (zero) values).

Multilocus correlated-matings model

Besides the selfing rate, an aspect of mating system is the degree that siblings share the same male parent. This is the outcome of pollination syndrome or population structure, and has implications for the effectiveness of certain types of natural selection. To characterize this sharing of

parentage, Ritland (1989) introduced the 'correlated-matings model'. This model treats a pair of progeny as the unit of observation. For mixed-mating populations, it has two parameters: the 'correlation of paternity', r_p , which is the probability that the two siblings are outcrossed full-sibs, and the 'correlation of selfing', r_s , between two members of a family (or the normalized variance of selfing rate among families).

In Ritland (1989), the probabilities of progeny under the correlated-matings model were defined in terms of a single, diallelic locus. Here, multiallelic, multilocus probabilities are given. Allowing for any number of alleles allows markers with greater statistical power (eg, microsatellites). More interestingly, comparison of single versus multilocus estimates of correlated matings (and not just selfing), provides a new measure of population substructure.

Let the two members of the progeny pair be $A_k A_l$ and $A_m A_n$, and the maternally-derived alleles be A_k and A_m . The maternal parent alleles are A_i and A_j (k equals either i or j , as does l). A pair of progeny can be: (1) both selfed ('ss'); (2) the first selfed and the second outcrossed ('st'); (3) the first outcrossed and the second selfed ('ts'); (4) both outcrossed to the same male ('t'); or (5) both outcrossed to different males ('tt'). Denote the probabilities of progeny pairs conditioned upon these five possibilities as $P_{kl, mn}^{ij, ss}$, $P_{kl, mn}^{ij, st}$, $P_{kl, mn}^{ij, ts}$, $P_{kl, mn}^{ij, t}$ and $P_{kl, mn}^{ij, tt}$, respectively.

Except for case (4), the probabilities of progeny pairs are simple extensions of equations (2) or (3):

$$\begin{aligned} P_{kl, mn}^{ij, ss} &= P_{kl}^{ij, s} P_{mn}^{ij, s} \\ P_{kl, mn}^{ij, st} &= P_{kl}^{ij, s} P_{mn}^{ij, t} \\ P_{kl, mn}^{ij, ts} &= P_{kl}^{ij, t} P_{mn}^{ij, s} \\ P_{kl, mn}^{ij, tt} &= P_{kl}^{ij, t} P_{mn}^{ij, t} \end{aligned} \quad (5)$$

where for the gametophyte-known case, the P 's are defined in equation (2), or for the gametophyte-unknown case, the P 's are defined in equation (3).

When outcrosses occur twice to the same male, the inbreeding coefficient of the male, termed f , enters the progeny probability. Noting that the correlation between two gametes from the same male is $(1+f)/2$, the expected frequency of the male-derived gamete pair (n, l) can be written as $p_{nl} = \left\{ (2-\delta_{nl})p_n p_l \left(\frac{1-f}{2} \right) + p_n \delta_{nl} \left(\frac{1+f}{2} \right) \right\}$. Like the classical assumption of constant selfing rates made by the mixed-mating model, this expectation assumes that males are equally inbred.

Now, for case (4), the probability of the progeny genotypes is, for the gametophyte-known case:

$$P_{kl, mn}^{ij, t} = D_k^{ij} D_m^{ij} p_{nl} \quad (6a)$$

or, for the gametophyte-unknown case:

$$\begin{aligned} P_{kl, mn}^{ij, t} &= \frac{1}{4} (2 - \delta_{kl}) (2 - \delta_{mn}) \\ &\quad (D_k^{ij} D_m^{ij} p_{nl} + D_k^{ij} D_n^{ij} p_{lm} + D_l^{ij} D_m^{ij} p_{kn} + D_l^{ij} D_n^{ij} p_{km}) \end{aligned} \quad (6b)$$

The multilocus probability of an offspring of mother ij is then:

$$\begin{aligned} &(s^2(1 - r_s) + sr_s) \prod_{\text{loci}} P_{kl}^{ij, ss} + st(1 - r_s) \\ &\quad \left(\prod_{\text{loci}} P_{kl}^{ij, st} + \prod_{\text{loci}} P_{kl}^{ij, ts} \right) + (t^2(1 - r_s) + tr_s) \end{aligned} \quad (7)$$

$$\left(r_p \prod_{\text{loci}} P_{kl}^{ij, t} + (1 - r_p) \prod_{\text{loci}} P_{kl}^{ij, tt} \right)$$

where r_s is the correlation of selfing and r_p is the correlation of paternity. This probability is then used in the likelihood equation for estimating correlated matings, as described in Ritland (1989).

Statistical properties: A FORTRAN 90 program was written to compute the information matrix for joint estimation of t , r_s , r_p and f (paternal inbreeding coefficient). Equations for the matrix of the second derivatives of the log-likelihood were found with an equation solver and imported directly into code. The expected value of the matrix was then found by randomly generating parents and pairs of progeny, using predefined gene frequencies and parameter values, and computing the second derivative of the likelihood of that observation, given the true parameter values. Ten thousand replications were performed, then the matrix inverted to give the variance-covariance matrix of estimates, on a per-pair basis.

The matrix was found to be singular for the case of one locus, meaning that the four mating system parameters (t , r_s , r_p and f) are not jointly estimable at a single-locus. Use of two or more loci resulted in non-singular matrices, showing that joint estimation of correlated-mating system parameters inherently requires a multilocus approach. In the single-locus model of Ritland (1989), f was not estimated and assumed equal to the maternal parent value.

Figure 1 shows the results for two to six gene loci, for the case of a two-allele locus and four-allele locus (for $t = r_p = 0.5$, $r_s = 0.25$ and $f = 0.2$). Additional loci decrease the variance of correlated-mating estimates, in the same manner as additional loci lower the variance of outcrossing estimates (Ritland and Jain, 1981). However, while assay of megagametophytes also lowered the standard error, the reduction is low and does not warrant the extra effort of megagametophyte assay.

Multilocus-based measures of biparental inbreeding

Biparental inbreeding, or mating between relatives, causes apparent selfing or increased homozygosity, relative to random mating. When true selfing is also present, the difference between multilocus and single-locus estimates of outcrossing, $(t_m - t_s)$, is often used to characterize the level of biparental inbreeding. This difference is positive because single-locus estimates includes all apparent selfing due to biparental inbreeding, whereas multilocus estimates exclude much of the apparent selfing due to biparental inbreeding (Shaw *et al*, 1980), as an observed outcross at any locus overrides the apparent selfing manifested at other loci. However, this difference is always an underestimate, as it depends upon the number of loci used, with more loci giving values closer to the true difference. A measure independent of the number of loci used is desirable.

An approximate measure of the fraction of apparent selfing due to biparental inbreeding, $(1 - r_s)$: When biparental inbreeding occurs, the multilocus segregation pattern of progeny is different from that expected under pure mixed selfing and outcrossing. 'Effective selfing' occurs at some loci and 'apparent outcrossing' at other loci (Ritland, 1984).

Figure 1 Expected standard errors, per pair of progeny, for estimates under the multilocus correlated-matings model, as a function of number of loci.

To derive a statistic that can be compared among studies, as a first approximation, consider two loci. At these loci, define the probability that both loci are selfed as s_{20} , the probability that one is selfed and the other outcrossed as s_{11} , and the probability that both loci are outcrossed as s_{02} . The selfing rates can be parameterized in terms of the single-locus selfing rate, s_s , and the correlation of selfing among loci, r_s :

$$\begin{aligned} s_{20} &= s_s^2 (1 - r_s) + s_s r_s \\ s_{11} &= 2s_s(1 - s_s) (1 - r_s) \\ s_{02} &= 1 - s_{11} - s_{20} = (1 - s_s)^2 (1 - r_s) \\ &\quad + (1 - s_s) (1 - r_s) \end{aligned} \quad (8)$$

These three probabilities correspond to loci being both selfed, one selfed/one outcrossed, and both outcrossed. The correlation of selfing among loci, r_s , lies between -1 and 1 , with values less than zero not possible under any realistic scenarios. Uniparental inbreeding results in complete correlation of selfing among loci, while biparental inbreeding results in a looser correlation. This suggests that r_s might be used to measure biparental inbreeding in some way.

Using gene-identity coefficients and the definition of 'effective selfing' (Ritland, 1984), the correlation of selfing among loci can be found for any mixture of matings. In outbred populations, the single-locus apparent selfing (E) caused by biparental inbreeding is twice the coefficient of relationship between the relatives ($E = \frac{1}{2}$ for full-sibs,

$E = \frac{1}{4}$ for half-sibs). For two loci, the apparent selfing rate is E^2 , and a given type of mating class, there is no among-locus correlation of selfing. However, when there is a mixture of matings including biparental inbreeding, the correlation takes an intermediate value. For example, in a population of 25% full-sib mating and 75% random outcrossing, $r_s = (0.25 \times 0.5^2 - 0.125^2)/(0.125 \times 0.875) \cong 0.43$.

Figure 2a plots r_s under a mixture of biparental inbreeding and random outcrossing, over a range of possible values for strength of biparental inbreeding (E) and the frequency of biparental inbreeding (s_b). It shows that under this scenario, the among-locus correlation of selfing r_s ranges from 0.4 to 1.0.

If there is a mixture of uniparental inbreeding (at rate s_u) and uniform biparental inbreeding (at rate s_b), the two locus and single-locus selfing rates are:

$$\begin{aligned} s_{20} &= s_b E_b^2 + s_u \\ s_s &= s_b E_b + s_u \end{aligned} \quad (9)$$

and equating s_2 in equation (9) with s_2 in equation (8) and solving for r yields the approximation:

Figure 2 (a) The expected among-locus correlation of selfing, as a function of the fraction matings that constitute biparental inbreeding (s_b), and the strength of biparental inbreeding (E). (b) The accuracy of r_s as a measure of the fraction of inbreeding due to uniparental inbreeding, as a function of the same two parameters.

$$r_s \cong \frac{s_u}{s_u + s_b E} \quad (10)$$

(this omits terms of order s^2 and E^2). Thus, for lower levels of selfing ($s < 0.2$), r_s is about the fraction of inbreeding due to uniparental inbreeding, and $(1 - r_s)$ is about the fraction of inbreeding due to biparental inbreeding.

This approximation appears to be quite good over a range of values, as indicated in Figure 2b, which plots the ratio of estimated to actual values assuming a selfing rate of 0.2 (lesser values give an even better fit). This relative bias is +5% to -15% over a range of reasonable values for s_b and E .

For estimation of the correlation of selfing among loci, we can use the notation of equation (4). With this, the likelihood of two-locus data is the mixture:

$$s_{20} P_{kl,1}^{ij,1,s} P_{kl,2}^{ij,2,s} + s_{11} (P_{kl,1}^{ij,1,s} P_{kl,2}^{ij,2,t} + P_{kl,1}^{ij,1,t} P_{kl,2}^{ij,2,s})/2 + s_{02} P_{kl,1}^{ij,1,t} P_{kl,2}^{ij,2,t}, \quad (11)$$

where the two loci are indicated by 1 and 2. If there are n loci, all $n(n-1)/2$ pairings enter as products of a pseudo-likelihood function.

Three-locus selfing model: With two loci, one cannot disentangle the two components of effective selfing, the frequency *vs* strength of biparental inbreeding. To disentangle these, at the simplest, it can be assumed that there is one type of biparental mating with strength E_b and frequency s_b , and that uniparental inbreeding also occurs at frequency s_u (the effective selfing rate given uniparental inbreeding, ' E_u ', equals one, so is omitted from the following). Suppose selfing rates are separately measured using single loci, pairs of loci and triplets of loci. Denoting these as s_1 , s_2 and s_3 respectively, their expectations are:

$$\begin{aligned} s_1 &= s_b E_b + s_u \\ s_2 &= s_b E_b^2 + s_u \\ s_3 &= s_b E_b^3 + s_u \end{aligned} \quad (12)$$

From these three expectations, joint estimators for the biparental inbreeding rate, and biparental inbreeding intensity, and the uniparental selfing rate can be obtained:

$$\begin{aligned} \hat{E}_b &= \frac{s_2 - s_3}{s_1 - s_2} \\ \hat{s}_u &= \frac{s_1 s_3 - s_2^2}{s_1 - 2s_2 + s_3} \\ \hat{s}_b &= \frac{s_1 - \hat{s}_u}{\hat{E}_b} \end{aligned} \quad (13)$$

With additional loci, more types of biparental inbreeding can be identified. For example, five loci possess the information to estimate two types of biparental inbreeding.

A precise multilocus description of biparental inbreeding would require n rates of selfing, s_i , $i = 1 \dots n$, where s_i is the frequency of selfing at i loci and outcrossing at the remaining $n - i$ loci. The single-locus selfing rate (s_s)

is the weighted average, $s_s = \left(\frac{1}{n}\right) \sum_i s_i$. While estimation

of all n selfing rates is possible, the statistical power to estimate all these parameters will be low, and furthermore, we still are left with measures of biparental

inbreeding that are not comparable among studies with different numbers of loci.

MME estimators for individual and family selfing rates

Many studies have estimated selfing rates at the family level (Ivey and Wyatt, 1999). However, when the maximum likelihood method is used, upward bias of outcrossing estimates occur (Ivey and Wyatt, 1999), particularly for smaller family sizes. To address this and other problems with the likelihood approach, Cruzan and Arnold (1994) developed a 'method of moments' estimator for selfing rate, which was based upon the Shaw *et al* (1980) multilocus estimator. Here, a similar, but more statistically efficient, estimator for family selfing rates is introduced. These estimators are developed using the same approach that gave estimators of pairwise relatedness (Ritland, 1996).

Single-locus estimators: The efficient single-locus 'method of moments' (MME) for selfing is based upon an indicator variable for each possible selfed genotype, conditioned upon parent genotype. These indicator variables are denoted by ' I ' and are defined as follows. If the parent is genotype $A_i A_i$, there is just one indicator variable termed ' I_{ii} ', defined such that if the progeny genotype is $A_i A_i$ then $I_{ii} = 1$; otherwise $I_{ii} = 0$. If the parent is genotype $A_i A_j$, there are three variables I_{ii} , I_{ij} , and I_{jj} , defined such that (a) if the progeny is $A_i A_i$, then $I_{ii} = 1$, $I_{ij} = 0$ and $I_{jj} = 0$; (b) if the progeny is $A_i A_j$, then $I_{ii} = 0$, $I_{ij} = 1$ and $I_{jj} = 0$; and (c) if the progeny is $A_j A_j$, then $I_{ii} = 0$, $I_{ij} = 0$ and $I_{jj} = 1$.

If the parent is genotype $A_i A_i$, the expectation of I_{ii} is $s + (1 - s)p_i$. Equating the observation of I_i to its expectation gives the MME for the homozygous parent:

$$\hat{s}_{ii} = \frac{I_{ii} - p_i}{1 - p_i}, \text{Var}(\hat{s}_{ii}) = \frac{\text{Var}(I_{ii})}{(1 - p_i)^2} = \frac{p_i}{(1 - p_i)} \quad (14)$$

assuming $s \cong 0$ so $\text{Var}(I_{ii}) = p_i(1 - p_i)$.

If the parent is genotype $A_i A_j$, the estimate of selfing is derived as follows. The values of I_{ii} , I_{ij} , and I_{jj} are equated to their expectations, resulting in three estimators. For example, the expected value of I_{ii} is $s/4 + tp_i/2$, and solving for s gives $\hat{s} = (4I_{ii} - 2p_i)/(1 - 2p_i)$. The estimate of selfing for parent $A_i A_j$ is a weighted summation of these three estimators, with the optimal weights determined by the variance-covariance matrix of estimates, and the optimum defined as that which minimizes $\text{Var}(\hat{s}_{ij})$; details of this derivation are too complex to present here, but follow the procedure used in Ritland (1996) for pairwise relatedness. The optimal estimator for a heterozygous parent was found with an analytical equation solver to be:

$$\begin{aligned} \hat{s}_{ij} &= \frac{4(I_{ii}p_j + I_{ij}p_i)(p_i + p_j) + 8I_{jj}p_i p_j - 8p_i p_j(p_i + p_j)}{(p_i - p_j)^2 + 8p_i p_j(1 - p_i - p_j)} \\ \text{Var}(\hat{s}_{ij}) &= \frac{8p_i p_j(p_i + p_j)}{(p_i - p_j)^2 + 8p_i p_j(1 - p_i - p_j)} \end{aligned} \quad (15)$$

Across loci, the above variances are used to obtain a weighted average of single-locus estimates across loci (each weight is the inverse of the variance). The family estimate of selfing is the weighted average of individual estimates, where the weight applied to a given individual is the average reciprocal variance across loci.

These formulae also illustrate the dependency of esti-

mation variance upon parent genotype and gene frequencies. The estimate for a homozygous parent is best when the parent is a rare genotype; estimates for heterozygote parents have much more uncertainty, and are particularly bad for diallelic loci with nearly equal gene frequencies.

These estimators, being conditioned on parent genotype, are similar to the 'regression' estimators of Lynch and Ritland (1999), where the statistical model involves the probability of a second individual, conditioned upon the first. By contrast, in Ritland (1996) the joint probabilities of both individuals were used in the statistical model.

Multilocus estimator: A multilocus MME for individual selfing rate involves a weighted average of selfing estimates across *all possible* selfed progeny genotypes. Using equation (4), each possible selfed progeny genotype gives an estimate and variance of:

$$\hat{s}_m = \frac{I_m - P_m^s}{P_m^s - P_m^t}, \text{Var}(\hat{s}_m) = \frac{P_m^t(1 - P_m^t)}{(P_m^s - P_m^t)^2} \quad (16)$$

where '*m*' is the *m*-th possible selfed multilocus progeny genotype of the parent, $P_m^s = \prod_{\text{loci}} P_{kl(m)}^s$, $P_m^t = \prod_{\text{loci}} P_{kl(m)}^t$, and $I_m = 1$ if genotype *m* is the observed progeny genotype, otherwise it is zero.

For *n* loci of which *p* are heterozygous in the parent, there are $2^{n-p}3^p$ possible selfed progeny genotypes. This can be a great number, and it becomes impractical to invert the variance-covariance matrix for obtaining an optimal weighted estimate. As an approximation, covariances can be neglected, and the multilocus estimate is simply the weighted average across all $\hat{s}_m$ estimates (weights are proportional to the inverse of each variance). The family estimate of selfing is the weighted average of individual estimates, where the weight applied to a given individual is the average reciprocal variance across loci.

A FORTRAN 90 computer program was written to simulate data and estimate individual selfing rates using these formula, and using Cruzan's method. The latter method (Cruzan and Arnold, 1994) is quite similar, the difference being that it groups together all three possible selfed genotypes of a heterozygous mother as the same observation. In the simulation, pollen gene frequencies were assumed to be known, eg, taken from a larger population sample, data were generated for cases of one to six loci and two or four alleles per locus, true outcrossing rate was 90%, and each combination was replicated 100 000 times.

The results (Figure 3) show that additional loci always decrease the variance of individual correlated-mating estimates, as do additional alleles per locus. The multilocus estimate is also considerably more accurate than the mean (across loci) of single-locus estimates. The Cruzan method cannot be evaluated for diallelic loci because a heterozygous mother is completely uninformative (infinite variance). For multiallelic loci, the Cruzan method is just slightly less efficient (5–10% greater standard error) than the full multilocus method (equation 16). Also, the Cruzan method makes no prior assumption of outcrossing rate, and its formula is considerably simpler. For these reasons, the Cruzan method is likely a better one to use, except when loci are all diallelic (this method is incorporated into MLTR).

Figure 3 Expected standard errors of individual estimates of selfing, as a function of number of loci used.

It has been assumed that pollen gene frequency is known or separately estimated. If pollen gene frequencies vary among plants, selfing rate estimates may be biased. However, at least two loci are needed to obtain the degrees of freedom for joint estimation of selfing and pollen pool. Issues regarding joint estimation of selfing and pollen gene frequency for individual families are discussed in Ivey and Wyatt (1999).

MLTR for Windows

The above improvements in the estimation of mating system parameters have been implemented in a complete rewrite and upgrade of 'MLT' (Ritland, 1990). The new program is termed 'MLTR' (for multilocus t and r), is written in Compaq Visual Fortran 95, and runs under Windows 95 and later operating systems. Run-time array dimensioning allows any size of sample (eg loci, alleles, families) to be used. Letters or numbers can be used for allele designations, and the megagametophyte may be either known or unknown. Additional features include: grouping of individuals for mating system estimation (gene frequencies assumed constant among groups), selection between the EM *vs* Newton-Raphson methods for numerical solution of the likelihood, and bootstrapping either within families or among families for determining errors of estimates. A complete description of capabilities and the formatting of data are available in the documentation. The program is currently available at: <http://www.genetics.forestry.ubc.ca/ritland/programs>.

One new feature is the option for 'Monte-Carlo' estimation of parentage. This is a modification of the Brown and Allard (1970) method for inferring the maternal parent, which incorporated the probability of all possible parents into the single-locus estimation procedure, and is preferable over using the 'most-likely' parent when progeny array size is smaller (<10). However, with multilocus procedures, there are too many possible multilocus parent genotypes to enumerate using the Brown and Allard method. Instead, we can emulate the Brown and Allard method by choosing the parentage for each locus at random, in proportion to the probability of parentage. With each bootstrap iteration, the selection is performed again. This Monte-Carlo method not only confers the desired properties of the Brown and Allard approach, but also incorporates the uncertainty about inferred parentage into the bootstrap estimate of statistical error. However, the original (non-bootstrapped) estimate is not unique; instead, the mating system estimates are the means of the bootstraps.

Discussion

This paper provides advances in four facets of mating system estimation: (1) new formulas for probabilities of progeny, allowing for any number of alleles; (2) a multilocus correlated-matings model; (3) new measures of biparental inbreeding; and (4) MME estimators for individual selfing rate. While each might be the topic of a separate paper, they share the common theme of being multilocus analyses of mating systems. As well, a single dataset is amenable to analyses by all these methods, which are implemented by the program, MLTR.

The significance of these advances is: (1) the formulae allow increased ease in programming, and show the relationship between the 'conifer' (gametophyte assayed) and 'angiosperm' models; (2) there is greater statistical power for correlated-matings estimates, a new measure of population substructure (comparison of single versus multilocus estimates of correlated-matings), and the power (albeit weak) to estimate paternal inbreeding coefficients from progeny arrays (when paternity is correlated); (3) nearly unbiased estimation of biparental inbreeding, as measured by the correlation of selfing among loci; and (4) unbiased estimators of individual and family selfing rate (when pollen gene frequency is not jointly estimated), suitable for fine-scale studies of mating systems.

The availability of new types of marker data, as well as the diversification of research into the ecology and evolution of plant mating systems, warrants this development of more sophisticated estimation methods. These methods generally require more informative data, either in terms of more loci or more alleles per locus, as finer features of the data are used. However, in many cases, isozyme data can still be adequate, provided that sample sizes are increased. Also, markers exhibiting dominance, such as AFLPs and RAPDs, can also be used in these models, as incorporating dominance into the probabilities is straightforward: the probability of a progeny phenotype equals the sum of the possible progeny genotypes underlying the phenotype. Dominant loci are allowed in MLTR.

Some of these facets of the mating system can be examined with alternative models. The correlated-matings

model is just one way to characterize paternity, as paternity analysis has attracted much interest. Wilson (1981) presented a statistical model for estimating the number of sires in a female brood, but with reference to animal populations. In plants, this approach was subsequently adopted (Ellstrand, 1984). Exclusion methods can identify specific male parents (Meagher, 1986), and fractional paternity assignment can distinguish among non-excluded parents (Devlin *et al.*, 1988). However, the power to identify parentage depends upon how informative (variable) the markers are, making comparisons between studies difficult.

Microsatellites can be sufficiently informative to even infer both parents of an individual (Gerber *et al.*, 2000). Experimental populations with mutually unique multilocus combinations of homozygous genotypes allow unambiguous assignment of paternity (Karron *et al.*, 1997). Alternatively, 'parametric' models can be used, as in the correlated-matings model (Ritland, 1989). The mixed-mating model can be modified to allow progeny arrays sired by one male parent (Schoen and Clegg, 1984). Also, an analogue of the correlation of paternity can be used to estimate average pollination distance (Austerlitz and Smouse, 2001).

There has been less interest in alternative measures of biparental inbreeding. The difference between multilocus versus single-locus outcrossing rate, $t_m - t_s$, is almost universally used. A measure of biparental inbreeding based upon the relationships between fixation indices of selfed and outcrossed progeny was proposed (Waller and Knight, 1989, equation 7), but this estimator essentially equals $(t_m - t_s)/t_m$, except that t_s is estimated using the fixation index of progeny.

A major problem with using the difference, $t_m - t_s$, is that it depends on the number of loci used, with more loci giving a larger difference due to larger t_m . The 'among-locus correlation of selfing', introduced in this paper, does not suffer this bias, and also directly approximates the fraction of selfing due to uniparental selfing. However, there are still cases when simple comparisons of selfing rates are useful, for example, when comparing populations of the same species for $t_m - t_s$ using the same set of loci (Lu, 2000). Also, the apparent single-locus selfing found in species with self-incompatibility such as sunflower (Ellstrand *et al.*, 1978) should be an unbiased estimator of biparental inbreeding, regardless of the number of loci used, as uniparental inbreeding cannot occur.

Individual and family estimates of selfing are useful in fine-scale studies of mating systems. However, such estimates suffer large statistical variance due to inherent small sample size, even with hypervariable data such as microsatellites. In fact, Ivey and Wyatt (1999) conclude that current family estimators are inadequate and that new procedures are needed. For individual or family-level estimates to be useful, the object of study should transcend these levels, and involve groups of such estimates or correlates of such estimates with other data. For example, individual selfing rates might be correlated with stigma-anther separation, and the object of estimation being the slope of the regression. Other recent advances in the marker-based approaches in ecology and evolution (Cruzan, 1998) are creating opportunities for the use of these fine-scale measures.

Acknowledgements

I thank Subodh K Jain for his continual encouragement; other colleagues, especially Spencer Barrett, for furthering my interests in mating system estimation, and the reviewers for their comments and corrections. This research was supported by grants from NSERC of Canada.

References

- Austerlitz F, Smouse PE (2001). Two-generation analysis of pollen flow across a landscape. II. Relation between Φ_{it} , pollen dispersal and interfemale distance. *Genetics* **157**: 851–857.
- Barrett SCH, Eckert CG (1990). Variation and evolution of mating systems in seed plants. In: Kawano S (ed). *Biological Approaches and Evolutionary Trends in Plants*. Academic Press: Tokyo, pp 229–254.
- Brown AHD, Allard RW (1970). Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. *Genetics* **66**: 133–145.
- Cheliak WM, Morgan K, Strobeck C, Yeh FCH, Dancik BP (1983). Estimation of mating system parameters in plant populations using the EM algorithm. *Theor Appl Genet* **65**: 157–161.
- Clegg M (1980). Measuring plant mating systems. *Bioscience* **30**: 814–818.
- Cruzan MB (1998). Genetic markers in plant evolutionary ecology. *Ecology* **79**: 400–412.
- Cruzan MB, Arnold ML (1994). Assortative mating and natural selection in hybridizing irises: effects of phenology and floral densities on family outcrossing rates. *Heredity* **72**: 95–105.
- Devlin B, Roeder K, Ellstrand NC (1988). Fraction paternity assignment: theoretical development and comparison to other methods. *Theor Appl Genet* **76**: 369–380.
- Ellstrand NC (1984). Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. *Am Nat* **123**: 810–828.
- Ellstrand NC, Torres AM, Levin DA (1978). Density and the rate of apparent outcrossing in *Helianthus annuus* (Asteraceae). *Syst Bot* **3**: 403–407.
- Fyfe JL, Bailey NTJ (1951). Plant breeding studies in leguminous forage crops. I. Natural crossbreeding in winter beans. *J Agric Sci* **41**: 371–378.
- Gerber S, Mariette S, Streiff R, Bodénès C, Kremer A (2000). Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Mol Ecol* **9**: 1037–1048.
- Hedrick P, Ritland K (1990). Gamete disequilibrium and multilocus estimation of selfing rates. *Heredity* **65**: 343–347.
- Ivey CT, Wyatt R (1999). Family outcrossing rates and neighborhood floral density in natural populations of swamp milkweed (*Asclepias incarnata*): potential statistical artifacts. *Theor Appl Genet* **98**: 1063–1071.
- Jones DF (1916). Natural cross-pollination in the tomato. *Science* **43**: 509–510.
- Karron JD, Jackson RT, Thumser NN, Schlicht SL (1997). Outcrossing rates of individual *Mimulus rigens* genets are correlated with anther-stigma separation. *Heredity* **79**: 365–370.
- Lu YQ (2000). Effects of density on mixed mating systems and reproduction in natural populations of *Impatiens capensis*. *Int J Pl Sci* **161**: 671–681.
- Lynch M, Ritland K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1753–1766.
- Meagher T (1986). Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely parents. *Am Nat* **128**: 199–215.
- Ritland K (1984). The effective proportion of self-fertilization with consanguineous matings in inbred populations. *Genetics* **106**: 139–152.
- Ritland K (1989). Correlated-matings in the partial selfer, *Mimulus guttatus*. *Evolution* **43**: 848–859.
- Ritland K (1990). A series of FORTRAN computer programs for estimating plant mating systems. *J Hered* **81**: 235–237.
- Ritland K (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res* **67**: 175–185.
- Ritland K (2000). Marker-inferred relatedness as a tool for detecting heritability in nature. *Mol Ecol* **9**: 1195–1204.
- Ritland K, Jain SK (1981). A model for the estimation of outcrossing rate and gene frequencies using n independent loci. *Heredity* **47**: 35–52.
- Ritland K, El-Kassaby YA (1985). The nature of inbreeding in a seed orchard of Douglas-fir as shown by an efficient multilocus model. *Theor Appl Genet* **71**: 375–384.
- Schoen DJ, Clegg MT (1984). Estimation of mating system parameters when outcrossing events are correlated. *Proc Nat Acad Sci USA* **11**: 5258–5262.
- Shaw DV, Kahler AL, Allard RW (1980). A multilocus estimator of mating system parameters in plant populations. *Proc Nat Acad Sci USA* **78**: 1298–1302.
- Smyth CA, Hamrick JL (1984). Variation in estimates of outcrossing in musk thistle populations. *J Hered* **75**: 303–307.
- Waller DM, Knight SE (1989). Genetic consequences of outcrossing in the cleistogamous annual, *Impatiens capensis*. 2. Outcrossing rates and genotypic correlations. *Evolution* **43**: 860–869.
- Wilson J (1981). Estimating the degree of polyandry in natural populations. *Evolution* **35**: 664–673.