*Heredity* (1981), **47**(1), 35-52 0018-067X/81/03150035$02.00



# A MODEL FOR THE ESTIMATION OF OUTCROSSING RATE AND GENE FREQUENCIES USING *n* INDEPENDENT LOCI

KERMIT RITLAND and SUBODH JAIN

*Department of Agronomy and Range Science, University of California, Davis, CA 95616*

Received 19.vi.80

SUMMARY

A mixed mating model for many unlinked loci is described. A procedure for estimation of the model parameters (outcrossing rate and gene frequencies), based on a multilocus maximum likelihood equation, is discussed and analyzed for bias, variance, and robustness. Genotypic data from families of known or unknown maternal parentage, or data from progenies of known maternal parentage, are used for estimation. The procedure is applicable to dominant or co-dominant Mendelian genes with two or three alleles per locus, and should be particularly useful in studies where the effort in scoring more loci is less than the effort in scoring more progeny. Variances of the multilocus estimates of outcrossing rate and pollen pool gene frequencies decrease when more loci are included in the estimation. Monte Carlo simulations showed the estimates to be unbiased when model assumptions are not violated, but the bias introduced by various violations is reduced when more loci are included in the estimate. Often the variance of a three or four locus estimate closely approaches the minimum variance possible (the variance of an estimate using infinitely many loci), setting a practical limit to the number of loci needed for a nearly minimum variance estimate. An example from some work on *Limnanthes* is presented to illustrate the use of multilocus model and its fit to data from natural populations.

## 1. INTRODUCTION

PLANT breeders, ecologists, and evolutionists have been generally interested in characterising the breeding system of plant species in terms of the mixed selfing and random mating model, wherein a certain proportion of zygotes are derived from self-fertilization and the remaining derived from random outcrossing, each generation. The breeding system fundamentally affects the genetic structure and dynamics of populations, and proper estimates of the outcrossing rates are often needed for evaluating various hypotheses concerning the effects of breeding systems, as well as for planning breeding programmes. In this paper, we use multilocus theory to develop a model and statistical procedure for the estimation of outcrossing rate and gene frequencies using the simultaneous segregation of alleles at many loci, and evaluate some statistical properties of this estimator.

The single locus version of estimation procedure developed in this paper traces back to the work of Fyfe and Bailey (1951) who also used the maximum likelihood method, based on fitting the observed proportions of genotypes descended from a known maternal genotype with the proportions expected under a mixed mating model. In this model, progenies of each maternal genotype represent a genetic array derived from ovules that outcross with probability $t$ to a pollen pool with gene frequency $p$, and self-fertilize with probability $(1-t)$. As both $t$ and $p$ are unknown, the two

estimates must be determined simultaneously. Genetic changes due to mutation or selection following fertilization, as well as any assortative mating or variability in pollen pool frequencies, are assumed to be absent. It is sufficient to bulk together progenies descended from all maternal parents of the same known genotype; Jain and Marshall (1967) described such an estimation procedure using the bulked seed lots of recessive and dominant classes.

Using electrophoresis, many segregating loci with co-dominant alleles can often be found in a population, and an increase in the power and versatility of estimation procedures is possible. However, Brown and Allard (1970) described the difficulty in electrophoretically assaying both the maternal parent and its progeny, since an allele is frequently expressed only at certain development stages, and seedling assays often disallow sampled plants to be saved for producing seed progeny. Therefore, Brown and Allard (1970) used the genotypic progeny array of each family to infer the maternal genotype of that family; this requires a sufficiently large family size to distinguish between the alternative segregation patterns of each possible maternal parent (see Brown, 1975). Since the segregation pattern of each maternal genotype depends on the outcrossing rate and pollen gene frequencies, and the solution of the likelihood equations for outcrossing rate ($\hat{t}$) and gene frequency ($\hat{p}$) depend on the inferred maternal genotypes, their procedure became an iterative two-step process, in which the most likely maternal parent for each family is inferred, thus allowing estimates of $t$ and $p$. Clegg, Kahler and Allard (1978) modified this procedure by including all the possible maternal genotypes of a family, weighted by their relative likelihoods, in the calculations (for a brief survey of various single locus models available for estimating outcrossing rate, see Jain 1979).

Vasek (1968) used progenies of genotypes recessive at two diallelic loci for an estimation of outcrossing; an appendix gave the 24 equations, worked out by Timothy Prout, describing the expected progeny ratios of the nine two-locus maternal genotypes. More recently, Brown, Zohary and Nevo (1978) used $n$-locus data for outcrossing estimation in a predominantly self-pollinated plant population. Their approach is an approximate multilocus extension of the method of Allard, Kahler and Weir (1972), an estimation procedure suitable only for organisms with low outcrossing rates. The estimator described here is a direct multilocus extension of the single locus estimation models using progeny arrays, utilizing a mathematical approach using matrices which we find more tractable even for the single locus case. We note that similar approaches to the multilocus problems in population genetics are currently of interest (Roux, 1974; Karlin and Liberman, 1979).

## 2. The probability model

The mixed mating model, with $n$ unlinked loci, uses a probability transition matrix whose $ij$th element is the probability of observing offspring genotype $i$ given maternal parent genotype $j$; these elements are a function of outcrossing rate and pollen gene frequencies. The matrix is postmultiplied by a column vector of maternal genotype frequencies to give offspring frequencies. The model is constructed here for the case of two co-dominant

alleles at each of the $n$ unlinked loci; other cases are noted in a later section of this paper.

We define

$$\boldsymbol{S}_k = \begin{bmatrix} 1 & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix}, \qquad \boldsymbol{T}_k = \begin{bmatrix} p_k & p_k/2 & 0 \\ q_k & \frac{1}{2} & p_k \\ 0 & q_k/2 & q_k \end{bmatrix},$$

$$f_k = \begin{bmatrix} f_{1,k} \\ f_{2,k} \\ f_{3,k} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{g}_k = \begin{bmatrix} g_{1,k} \\ g_{2,k} \\ g_{3,k} \end{bmatrix} = \begin{bmatrix} A_kA_k \\ A_ka_k \\ a_ka_k \end{bmatrix},$$

where $p_k$ and $q_k (= 1 - p_k)$ are allelic frequencies (alleles symbolized by $A_k$ and $a_k$) in the pollen pool, and $f_{1,k}$, $f_{2,k}$, and $f_{3,k}$ are genotypic frequencies, for the $k$th locus. The $ij$th element of $\boldsymbol{S}_k$ is the conditional probability of progeny $g_{i,k}$ given it is progeny from a self-fertilization of maternal parent $g_{j,k}$, and the $ij$th element of $\boldsymbol{T}_k$ is the conditional probability of progeny $g_{i,k}$ given it is an outcrossed progeny of maternal parent $g_{j,k}$. The vector $\boldsymbol{g}_k$ labels genotypes referred to by the corresponding elements of $\boldsymbol{S}_k$, $\boldsymbol{T}_k$, and $f_k$, and is not used in any computations.

We now use the Kronecker product of matrices, denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$ for the product of $\boldsymbol{A}$ with $\boldsymbol{B}$. This operation takes each element of $\boldsymbol{A}$ and scalar multiplies the entire $\boldsymbol{B}$ matrix, generating a new submatrix of the dimensions of $\boldsymbol{B}$ in place of each $\boldsymbol{A}$ element, resulting in a matrix of size $(ik \times jl)$ if $\boldsymbol{A}$ is of dimension $(i \times j)$ and $\boldsymbol{B}$ is of dimension $(k \times l)$, *viz.*,

$$A \otimes B = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_1, {}_{j}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_2, {}_{j}B \\ \vdots & & & \vdots \\ a_{k1}\boldsymbol{B} & a_{k2}\boldsymbol{B} & \cdots & a_{ij}\boldsymbol{B} \end{bmatrix},$$

The Kronecker product of two column vectors is identical to the Kronecker product of two single column matrices. If we let

$$\boldsymbol{S} = \boldsymbol{S}_1 \otimes \boldsymbol{S}_2 \otimes \cdots \boldsymbol{S}_n$$

$$\boldsymbol{T} = \boldsymbol{T}_1 \otimes \boldsymbol{T}_2 \otimes \cdots \boldsymbol{T}_n$$

then for $n$ unlinked loci and outcrossing rate $t$, the parent-offspring transition matrix $\boldsymbol{P}$ is

$$\boldsymbol{P} = (1 - t)\boldsymbol{S} + t\boldsymbol{T}.$$

Observe that $\boldsymbol{P}$ is a function of outcrossing rate and pollen gene frequencies, with column sums equal to unity, and that $\boldsymbol{P}$ can get large: for $n$ loci and two alleles, the matrix is of size $(3^n \times 3^n)$ and has $3^{2n}$ elements.

Let $\boldsymbol{m}$ be a vector of length $3^n$ which contains maternal genotypic frequencies, and whose $i$th element is the frequency of the multilocus genotype named by the $i$th term of the Kronecker "product"

$$\boldsymbol{g} = \boldsymbol{g}_1 \otimes \boldsymbol{g}_2 \otimes \cdots \boldsymbol{g}_n.$$

(This "product" associates names between loci.) Now it is apparent that, by referring to $\boldsymbol{g}$, the $j$th column of $\boldsymbol{P}$ is the probability distribution of

progeny from the multilocus maternal parent genotype named by the $j$th element of $\boldsymbol{g}$. Furthermore, the $ij$th element of $\boldsymbol{P}$ contains the probability of the progeny genotype named by the $i$th element of $\boldsymbol{g}$, given the maternal parent genotype named by the $j$th element of $\boldsymbol{g}$. Therefore, the probability of the progeny genotype named by the $i$th element of $\boldsymbol{g}$, regardless of parentage, is the $i$th element of the column vector resulting from the matrix-vector product $\boldsymbol{Pm}$. The estimation procedure in this paper is based on this model, which assumes no selection or mutation following fertilization, all genotypes outcross at the same rate to a homogeneous pollen pool, and alleles at different loci segregate independently.

We need an equation for the frequencies of multilocus genotypes in a mixed mating population at equilibrium (assuming no natural selection, mutation, or linkage between loci) because the inference of maternal parentage will partly depend on these frequencies. Analogous to Wright's equilibrium for a single locus under inbreeding, we define $\boldsymbol{f}$ as the column vector of inbreeding equilibrium genotypic frequencies (naming by the elements of $\boldsymbol{g}$ is now implicitly assumed), and $\boldsymbol{f}_o$ as the column vector of multilocus Hardy-Weinberg genotypic frequencies (where each locus is at Hardy-Weinberg equilibrium and no genotypic associations exist between loci). $\boldsymbol{f}_o$ is simply found by taking the Kronecker product of all single locus Hardy–Weinberg frequency vectors. To find $\boldsymbol{f}$, note that at equilibrium,

$$\boldsymbol{f} = \boldsymbol{Pf},$$

or

$$\boldsymbol{f} = (1-t)\boldsymbol{Sf} + t\boldsymbol{Tf}.$$

Then,

$$\boldsymbol{f} = (1-t)\boldsymbol{Sf} + t\boldsymbol{f}_o,$$

since one generation of random outcrossing restores multilocus Hardy–Weinberg equilibrium (no gametic disequilibrium is generated by the mixed mating model). Solving for $\boldsymbol{f}$ yields

$$\boldsymbol{f} = t(\boldsymbol{I} - (1-t)\boldsymbol{S})^{-1}\boldsymbol{f}_o,$$

where $\boldsymbol{I}$ is the identity matrix of suitable dimension. The equilibrium frequencies are thus obtained by inverting a matrix which depends only on the outcrossing rate, and multiplying the inverted matrix with a vector of Hardy–Weinberg frequencies and the outcrossing rate. The elements of the inverted matrix can actually be found by inverting a matrix of size $[(n+1)\times(n+1)]$, for $n$ loci, thus avoiding the prohibitive task of inverting $(\boldsymbol{I}-(1-t)\boldsymbol{S})$ for large $n$. This equilibrium inbreeding frequency vector $\boldsymbol{f}$ accounts for the occurrence of genotypic identity disequilibrium, or the concentration of heterozygosity in fewer individuals than expected, on the basis of single locus products, in mixed mating populations, an effect originally studied by Bennett and Binet (1956), and further analyzed by Weir and Cockerham (1973).

## 3. The estimation procedure

We are interested in estimating from the progeny data over $n$ loci the outcrossing rate $t$, a column vector $\boldsymbol{p}$ of pollen gene frequencies (of length

$n$), and if we infer maternal genotypes, the vector of maternal genotypic frequencies $\boldsymbol{m}$ (of length $3^n$). The estimation of these parameters uses a two-step maximum likelihood procedure as follows. The first step (to be skipped if maternal parentage is known) is to derive a matrix of probable maternal genotypes, based upon the data and prior estimates or guesses of $t$, $\boldsymbol{p}$ and $\boldsymbol{m}$. For $l$ families, the progeny data are placed in a matrix $\boldsymbol{D}$ of size $(3^n \times l)$, whose $ij$th element contains the observed integer number of progeny of genotype $i$ in the $j$th family. Let $\boldsymbol{M}$ be a matrix of size $(l \times 3^n)$, whose $ij$th element is the probability of maternal parentage of the $i$th family by the $j$th genotype.

Using Baye's theorem, the $ij$th element of $\boldsymbol{M}$ is estimated by raising each element in the $j$th column of $\boldsymbol{P}$ to the power of each corresponding element (*i.e.*, elements in the same position) in the $i$th column of $\boldsymbol{D}$, then taking the product of all these $3^n$ terms with the $j$th element in $\boldsymbol{m}$, and finally normalizing the product with respect to the entire row of $\boldsymbol{M}$. This element is then the probability that genotype $j$ is the true maternal parent of the $i$th family, given $\boldsymbol{Pm}$, and the data in the $i$th column of $\boldsymbol{D}$. An alternative (following Brown and Allard 1970) is to choose the most likely maternal parent as the only parent used in subsequent calculations; then the rows of $\boldsymbol{M}$ are set to zero except for the maximum value of each row, which is set to one. The matrix of observed parent–progeny transitions $\boldsymbol{X}$ is then

$$\boldsymbol{X} = \boldsymbol{DM},$$

whose $ij$th element contains the number (not necessarily integer-valued) of progeny of genotype $i$ descended from parent genotype $j$, in the entire population. If maternal parentage is known, or if the progenies are offspring of bulked seedlots (each lot is of uniform genotypic composition), the data are placed directly into $\boldsymbol{X}$ at this point.

We now define $\boldsymbol{A} \circ \boldsymbol{B}$, for $\boldsymbol{A}$ and $\boldsymbol{B}$ of identical dimensions, as the element by element product of $\boldsymbol{A}$ with $\boldsymbol{B}$ (often referred to as the Schur product), wherein each element of $\boldsymbol{A}$ is multiplied by the corresponding element in $\boldsymbol{B}$, resulting in a new matrix of the same dimension. We also define $\ln(\boldsymbol{A})$ as the natural logarithm of each element in $\boldsymbol{A}$ (resulting in a matrix of the same dimension), $\mathbf{1}$ as a column vector of ones of length $3^n$, and $\boldsymbol{A}^T$ as the transpose of $\boldsymbol{A}$.

The second step is to fit the observed transitions in $\boldsymbol{X}$ with those expected in $\boldsymbol{P}$ by maximizing the multilocus log likelihood equation

$$L(\boldsymbol{X}, \boldsymbol{P}) = \mathbf{1}^T(\boldsymbol{X} \circ \ln(\boldsymbol{P}))\mathbf{1}.$$

As $\boldsymbol{P}$ is a function of $t$ and $\boldsymbol{p}$, $\boldsymbol{P}$ is adjusted until $L(\boldsymbol{X}, \boldsymbol{P})$ attains the maximum value. The estimation is finished at this point if maternal parentage is known or bulks are used; otherwise several repetitions of the procedure are required to improve the prior estimates of $t$, $\boldsymbol{p}$ and $\boldsymbol{m}$.

To obtain a revised estimate of $\boldsymbol{m}$, we first estimate single locus maternal (ovule) allele frequencies ($o_k$) using the data in $\boldsymbol{M}$. For locus $k$,

$$\hat{o}_k = \mathbf{1}^T\boldsymbol{M}((111)\otimes(111)\otimes\cdots(1\tfrac{1}{2}0)\otimes\cdots(111))^T$$

where $\mathbf{1}$ is a column vector of length $l$, and the vector $(1\ \frac{1}{2}\ 0)$ is the $k$th 3-element vector in the above Kronecker product. Let $\hat{\boldsymbol{m}}_{o,k}$ be a three element vector of Hardy-Weinberg frequency estimates at the $k$th locus,

and $\hat{\boldsymbol{m}}_o$ be a $3^n$ element vector of multilocus Hardy–Weinberg frequency estimates. The Hardy–Weinberg estimates $\hat{o}_k^2$, $2\hat{o}_k(1-\hat{o}_k)$ and $(1-\hat{o}_k)^2$, are placed into $\hat{\boldsymbol{m}}_{o,k}$ for each locus $k$; and the Kronecker product over all $\hat{\boldsymbol{m}}_{o,k}$, $1 \leqq k \leqq n$, is taken to obtain $\hat{\boldsymbol{m}}_o$. Then an estimate of $\boldsymbol{m}$ is $\hat{\boldsymbol{m}} = \hat{t}(\boldsymbol{I}-(1-\hat{t})\boldsymbol{S})^{-1}\hat{\boldsymbol{m}}_o$ (as derived earlier). The estimates of single locus pollen gene frequencies are directly obtained when $L(\boldsymbol{X}, \boldsymbol{P})$ is maximized, for the elements of $\boldsymbol{T}$ (all of which contribute to $\boldsymbol{P}$) are taken to be the product, over all loci, of single locus pollen gene frequencies. Note, however, that these single locus gene and genotypic frequency estimates are a function of the data from all loci.

To maximize $t$ in $L(\boldsymbol{X}, \boldsymbol{P})$, we use the one-step iteration

$$t_{i+1} = \mathbf{1}^T(\boldsymbol{X} \circ t_i \boldsymbol{T}/\boldsymbol{P}k)\mathbf{1}$$

where $\boldsymbol{T}/\boldsymbol{P}$ is the element by element quotient (each element of $\boldsymbol{T}$ is divided by the corresponding element of $\boldsymbol{P}$ resulting in a matrix of identical dimensions), $k$ is sample size and $\mathbf{1}$ is a column vector of ones of length $3^n$; a similar expression for $\boldsymbol{p}$ was used. This method always, but sometimes slowly, converges for $t < 1$. Maximization was approximately achieved as judged by the printouts of the likelihood value at each iteration; convergence of this simple iteration occurs perhaps because the second derivatives of the log-likelihood function are negative or zero, except for $\partial^2 L(\boldsymbol{X}, \boldsymbol{P})/\partial t\, \partial p$ which approaches 0 as $n$ gets larger. Fewer iterations were needed when more loci were included, as the maximum becomes sharper peaked, but more iterations and occasionally uncertain convergence were typical when loci exhibited dominance or when simulated data (discussed later) were generated that involved departures (selection, assortative mating) from the assumptions.

With a large number of loci, the size of these matrices would exceed the computer memory, and if all possible multilocus parents were included, the computational time could become prohibitive. But it is sufficient to work only with those elements of $\boldsymbol{P}$ corresponding to the non-zero elements of $\boldsymbol{X}$; *i.e.*, we need to consider only the likelihood of each sampled genotype. If parentage is known, the number of elements of $\boldsymbol{P}$ used in computations are merely equal to or less than the number of multilocus genotypes sampled. Inferring maternal parentage is more difficult; the procedure we adopt is to first eliminate very unlikely parents by using single locus analysis, then examine the remaining possible maternal parents using multilocus analysis, and excluding from $\boldsymbol{M}$ the less probable multilocus genotypes if their likelihood was less than some fraction of the most likely multilocus genotype. An estimation program, written in FORTRAN for a PDP 11/34 minicomputer, which can accommodate up to 10 loci, 500 plants, two or three alleles per locus, and either dominant or codominant expression of alleles, is available from K. Ritland.

## 4. Generality of the model

The model presented here can be modified to incorporate some other modes of inheritance. $\boldsymbol{S}_k$ and $\boldsymbol{T}_k$ can be enlarged to accommodate three or more alleles at locus $k$. Dominance is incorporated at the $k$-th locus by adding together rows of the $\boldsymbol{S}_k$ matrix which correspond to the same offspring phenotype, and likewise adding together all corresponding rows

of the $\boldsymbol{T}_k$ matrix. Then the Kronecker products and computations proceed as before, with the observation that $\boldsymbol{P}$ and $\boldsymbol{X}$ now have fewer rows than columns. For data on triploid endosperm characters (when two identical alleles are derived from the maternal side, the third from the paternal) the diallelic transition matrix is of size $(4 \times 3)$. If genotypic differences in outcrossing rates are of interest, $\boldsymbol{P}$ can be expressed (using the element by element product of a matrix with a vector, where each element of the vector scalar multiples the corresponding column of the matrix) as $\boldsymbol{P} = (\boldsymbol{S} \circ (\boldsymbol{1} - \boldsymbol{t}) + \boldsymbol{T} \circ \boldsymbol{t})$ where $\boldsymbol{t}$ and $\boldsymbol{1}$ are now column vectors of length $3^n$, and each element of $\boldsymbol{t}$ is estimated.

A multilocus estimation model of Shaw, Kahler and Allard (1981) can be compared with this model. Briefly, the expected proportion of the data elements of $\boldsymbol{X}$ corresponding those elements of $\boldsymbol{P}$ containing *non-zero* contributions from both $\boldsymbol{S}$ and $\boldsymbol{T}$ is calculated from the single locus pollen gene frequency estimates (obtained using the model of Clegg *et al.*, 1978) and known maternal parent genotypes. This proportion (the non-discernable outcrosses) is multiplied by $t$, and added on to the proportion of the data containing directly observable outcrosses (the proportion of the elements of $\boldsymbol{X}$ corresponding the elements of $\boldsymbol{P}$ with a *zero* contribution from $\boldsymbol{S}$; this is when a homozygous mother has heterozygous progeny at some locus), to yield a multilocus estimate of $t$. This method does not make efficient use of the multilocus data for estimating $t$, particularly when there are few loci available or gene frequencies are extreme, but is notable in its simple and direct approach in using some properties of a multilocus inbreeding system.

## 5. Properties of the estimates

To characterize basic statistical properties some analytical results will be presented, but often properties had to be evaluated using Monte Carlo simulation. For simulation some hypothetical data sets were made up as follows: (i) outcrossing rate and gene frequencies were specified; (ii) for each family a maternal genotype was randomly chosen at each locus according to the equilibrium proportions under inbreeding; (iii) for each progeny in a family a self or outcross was randomly chosen according to their respective probabilities; and (iv) for each progeny, genotypes were chosen randomly at each locus according to the probabilities contained in the columns in $\boldsymbol{S}_k$ (if selfed) or $\boldsymbol{T}_k$ (if outcrossed), for $1 \leqq k \leqq n$. This describes the basic model with no selection or assortative mating; however, in order to study their effects on the estimates, modified data sets were generated as noted below. Once a complete data set was generated, estimates of $t$, $\boldsymbol{p}$ and $\boldsymbol{m}$ were obtained using the estimation procedure. The process was replicated many times (from 50 to 250) until the estimates gave confidence intervals indicating the bias or variance of the estimates. Ten families with ten individuals per family were used with input $p_k = 0{\cdot}5$, $t = 0{\cdot}5$, unless otherwise stated.

### (i) *Bias*

A deviation of the mean of many replicate estimates from the input value indicates a bias in the estimate. No significant deviations were found

when using the basic multilocus model, but when selection or assortative mating modified the data, the estimator displayed some interesting properties. In considering the effects of selection, it is important to note that selection for heterozygotes has two effects on the data. First, it raises the pre-selection effective outcrossing rate by the selective deaths of proportionally more selfed zygotes; this "post-selection outcrossing rate" is $t_s = (tw^T\boldsymbol{Tm})/(w^T\boldsymbol{Pm})$, where $w$ is a column vector of fitness coefficients. Second, selection raises the level of heterozygosity in both selfed and outcrossed progeny. To study how these two factors affect the multilocus outcrossing estimate, two sets of data were generated. First, homozygotes at one of the $n$ loci were assigned fitnesses of 0·667, relative to 1 for the heterozygote, whereas all other loci were neutral. Second, all loci were made heterotic, with homozygotes at all loci assigned relative fitnesses of 0·667. Results (fig. 1) show that the multilocus estimate asymptotically

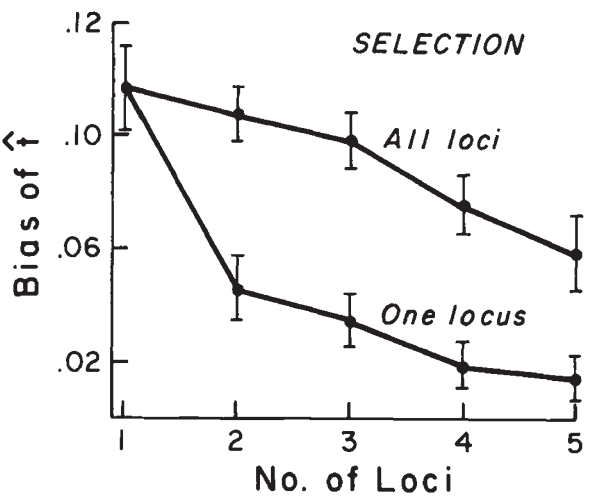


FIG. 1.—Deviation or bias of the multilocus estimate of outcrossing from the post-selection outcrossing rate $t_s$ with heterotic selection at all loci, and heterotic selection at just one locus (estimated by simulations described in text). 95 per cent confidence intervals are given in all figures.

approaches the post-selection outcrossing rate $t_s$ with added loci; the deviation or bias of the estimate from the post-selection outcrossing rate decreases slowly if all loci are under selection, and decreases quickly if only one locus is under selection. Similar results were obtained for the case of selection against heterozygotes.

To study the effects of non-random mating during outcrossing, the outcrossing of a homozygote at a locus to the identical pollen type was reduced by one-half. One set of simulations was run with one locus mating in this negative assortative mode, and remaining loci with random mating, while a second set of simulations were run with all loci under negative assortative mating. The results (fig. 2) indicate a pattern of bias reduction similar to the selection study (quick elimination with just one locus in violation, slow elimination with many loci in violation). A similar pattern

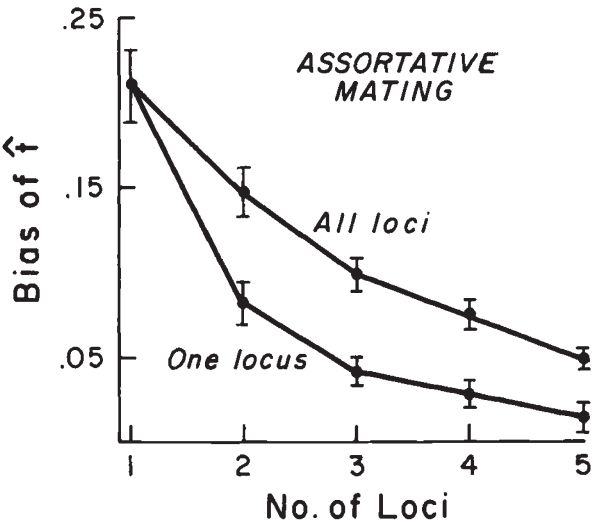


FIG. 2.—Deviation or bias of the multilocus estimate of outcrossing from the true outcrossing rate with negative assortative mating at all loci, and at one locus (estimated by simulations described in text).

of bias reduction was observed in a simulation of positive assortative mating, and when data were generated with positive assortative mating at some loci and negative assortative mating at other loci (or plus–minus selection effects) the biases were observed to cancel each other out. Shaw and Allard (1979) discuss how hypothesized differences in the bias of single vs. multilocus estimates can provide additional information concerning the breeding structure of a population. Our simulation results confirm that the bias of a multilocus estimate is reduced relative to a single locus estimate, but that elimination of bias with additional loci can be gradual in some cases.

To test for any deviations of data from the assumptions of the model (such as those mentioned above), the chi-square test for the goodness of fit of model expectations with the data, using element by element products and quotients, is given for $\boldsymbol{E} = k\boldsymbol{P} \circ \boldsymbol{m}$, by

$$\chi^2 = \mathbf{1}^T((\boldsymbol{X} - \boldsymbol{E}) \circ (\boldsymbol{X} - \boldsymbol{E})/\boldsymbol{E})\mathbf{1},$$

where $\mathbf{1}$ is a column vector of length $3^n$, with $7^n - 4\sum_{i=1}^{n} 7^{i-1} - (n+1)$ degrees of freedom (for $n$ diallelic loci). However, the number of categories (*e.g.*, non-zero elements of $\boldsymbol{P}$) increases by $7^n$, necessitating the lumping of categories for larger $n$, since the chi-square test is not valid if many cells have low expectations. But the categories must be lumped randomly (otherwise certain loci will be favoured), involving a rather difficult task. A less satisfactory but feasible test for goodness of fit of multilocus data, which we will use in our example, is to obtain an empirical distribution of the value of the log-likelihood function, using Monte Carlo simulation with the estimates as input parameters, and to reject the model if the log-likelihood of the estimates and data in question is less than, say, 95 per cent of the log-likelihood values based on simulated data sets.

(ii) *Variance*

Variances of the estimates $\hat{t}$, $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{m}}$ decrease, sometimes rapidly, when more loci are used, the rate of decrease depending on the true values of parameters, the number of progeny per family if maternal genotypes are inferred from progeny arrays, and the genotypic composition of the maternal population.

If maternal parentage is known and sample size is large, the lower bound for the variances and covariances of $\hat{t}$ and $\hat{\boldsymbol{p}}$ are found by inverting a symmetric information matrix. Using element by element products and quotients, element (1, 1) of the diallelic multilocus information matrix is given by the expected value of the second derivative with respect to $t$, viz.

$$-E(d^2L(\boldsymbol{X}, \boldsymbol{P})/dt^2) = \mathbf{1}^T((\boldsymbol{T}-\boldsymbol{S})\circ(\boldsymbol{T}-\boldsymbol{S})/\boldsymbol{P})\boldsymbol{m}$$

where $\mathbf{1}$ is a column vector of ones of length $3^n$.

Elements $(i+1, i+1)$ for $1 \leqq i \leqq n$, are obtained from the expected second derivatives with respect to $p$ at the $i$th locus,

$$-E(d^2L(\boldsymbol{X}, \boldsymbol{P})/dp_i^2) = \mathbf{1}^T((t^2\boldsymbol{T}'(i)\circ\boldsymbol{T}'(i)/\boldsymbol{P})\boldsymbol{m}$$

where $\boldsymbol{T}'(i)$ is identical to $\boldsymbol{T}$ except that the $i$th matrix ($\boldsymbol{T}_i$) in Kronecker product forming $\boldsymbol{T}'(i)$ is replaced by the matrix whose elements are derivatives, with respect to $p_i$, of each corresponding element of $\boldsymbol{T}_i$. Elements $(1, j+1)$ and $(j+1, 1)$, $1 \leqq j \leqq h$, are expected second derivatives with respect to $t$ and $p_j$,

$$-E(d^2L(\boldsymbol{X}, \boldsymbol{P})/dt\,dp_j) = \mathbf{1}^T(t\boldsymbol{T}'(j)\circ(\boldsymbol{T}-\boldsymbol{S})/\boldsymbol{P})\boldsymbol{m},$$

and elements $(i+1, j+1)$ and $(j+1, i+1)$, $1 \leqq i < j \leqq n$, are expected second derivatives with respect to $p_i$ and $p_j$,

$$-E(d^2L(\boldsymbol{X}, \boldsymbol{P})/dp_i\,dp_j) = \mathbf{1}^T(t^2\boldsymbol{T}'(i)\circ\boldsymbol{T}'(j)/\boldsymbol{P})\boldsymbol{m}.$$

Inverting this matrix gives variance of $\hat{t}$ per observation in element (1,1), the variances of $\hat{p}_i$ per observation on the remaining diagonal elements, and covariances on the off-diagonals.

Using this formula, a plot of the variance of $\hat{t}$ per observation from 1 to 5 loci for gene frequencies of 0·5 at inbreeding equilibrium (fig. 3) reveals the decrease in variance, which is especially pronounced for $t$ greater than 0·5 as some additional loci are used in the estimate. The variance per observation with infinitely many loci is $t(1-t)$, and even with few loci this limiting variance becomes the dominant component of the variance. However, our calculations of variances when gene frequencies are more extreme or when loci exhibit dominance, show a less rapid approach to the limiting value as more loci are added, indicating that relatively more loci are needed in these cases for a satisfactory estimate of $t$.

The variances of $\hat{p}_i$ found by this formula, for gene frequencies of 0·5 and $t$ of 0·5, (at inbreeding equilibrium) are plotted in fig. 4 for cases of both dominant and codominant loci (the variance is symmetric about $p = 0{\cdot}5$ for codominant loci). The decrease in variance is not as great as in fig. 3, but reduction is still pronounced, especially with dominance and high $p$. When dominant gene frequencies were near one, the variance of the single locus estimate $\hat{p}_i$ (which approached infinity as $p$ approached 1) was reduced to near 6 with the addition of one locus. The theoretical limit of the variance

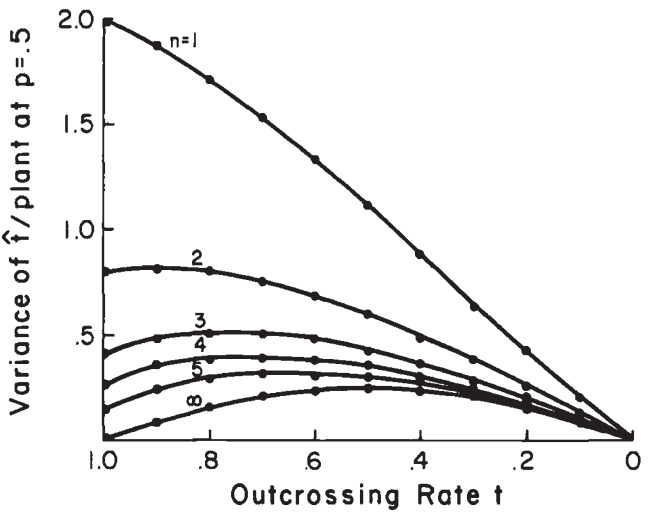


FIG. 3.—Variance of the estimate of outcrossing rate $\hat{t}$ per observation, as a function of the true outcrossing rate $t$ and number of loci $n$, for gene frequencies of 0·5 (determined by inverting the information matrix).

of pollen gene frequency estimate with infinite loci is not $p_k(1-p_k)/t$ since at that one locus case, heterozygotes in the progeny of heterozygous parents yield no information about the pollen gene frequencies at that locus regardless of the number of other loci used in the estimate. The limit for the co-dominant case is $p_k(1-p_k)/(t(1-m_{2,k}/2))$ for the $k$th locus, and for the

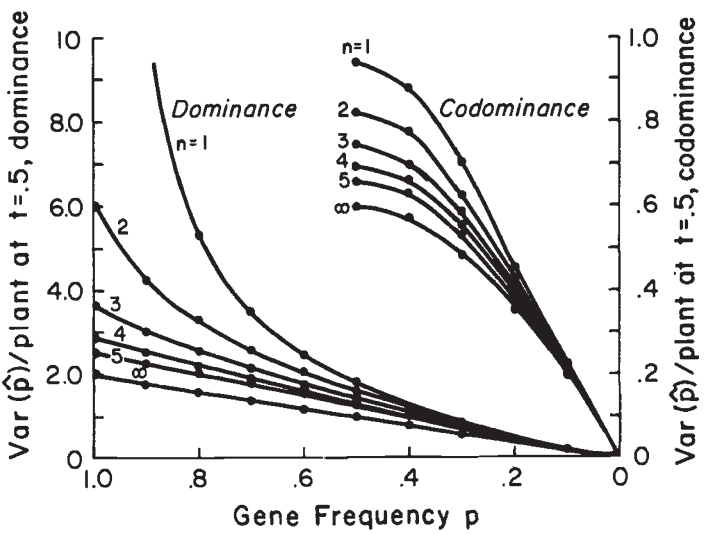


FIG. 4.—Variance of the estimate of pollen gene frequency $\hat{p}$ at a particular locus for $t = 0{\cdot}5$, as a function of pollen gene frequency $p$, number of loci $n$ (gene frequencies are identical over all loci) and mode of inheritance (determined by inverting information matrix). Variances are symmetric about $p = 0{\cdot}5$ for loci with codominance, so only right half is shown.

dominance case it is $p_k(1-p_k)/(t(1-m_{1,k}-m_{2,k}/2))$, where $m_{1,k}$, $m_{2,k}$ and $m_{3,k}$ are frequencies of maternal genotypes $A_kA_k$, $A_ka_k$, $a_ka_k$, respectively, at the $k$th locus.

The total variance of estimates would include effects due to a small total sample size, the sampling of families, misidentifications when inferring maternal genotypes, as well as numerical problems in maximizing the log-likelihood equation. Simulations, with parameters as described at the start of this section, were run to estimate the total variance for various values of $t$ and $n$ (note that sample size is 100). Simulation results (fig. 5) again demonstrate the decrease in variance of the estimates when more loci are used in the estimate, especially marked between one and two loci. When gene frequencies are intermediate, fewer loci are needed for species with low outcrossing rates to closely approach the theoretical limiting variance, whereas more loci (perhaps 4 or 5) are needed for species with

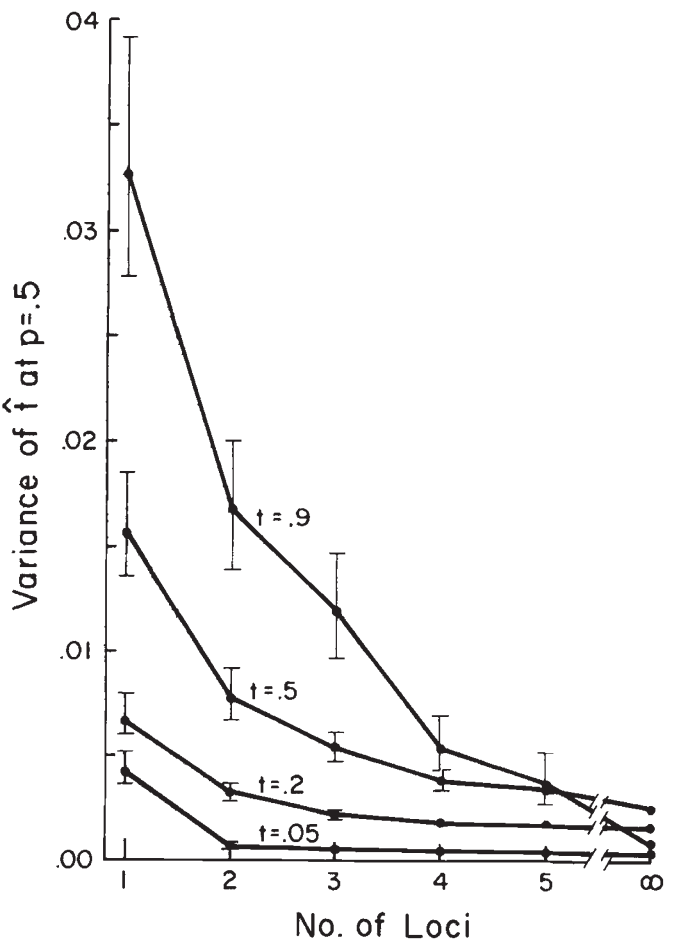


FIG. 5.—Results of simulations (discussed in text) showing the decrease in variance of the estimate of outcrossing rate $\hat{t}$, for gene frequencies of 0·5, when more loci are included in the estimate.

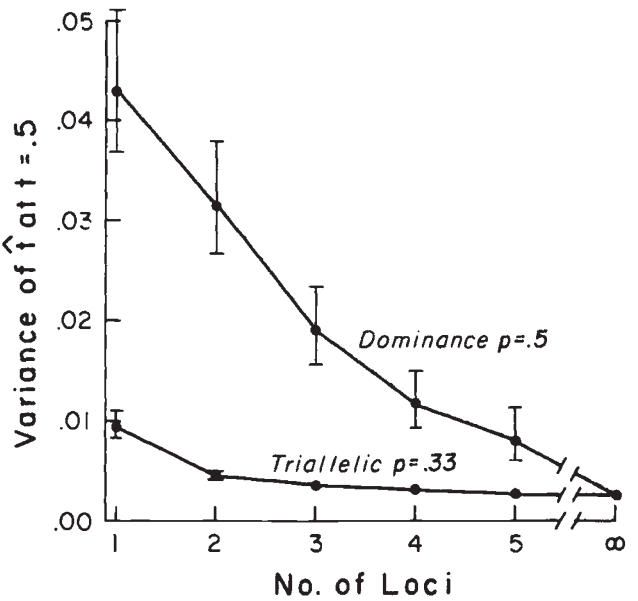


FIG. 6.—Simulation results demonstrating the decrease in variance of the estimate of outcrossing rate $\hat{t}$ when more loci are included in the estimate, for two modes of inheritance and outcrossing rate of 0·5. Upper line shows decrease when all loci exhibit dominance and gene frequencies are 0·5, while lower line shows decrease when all loci have 3 codominant alleles at frequencies of 0·333.

high outcrossing rates. The effects of dominant alleles and triallelic loci were also evaluated using simulation and the results (fig. 6) for $t = 0{\cdot}5$, show a continual decrease in variance when alleles are dominant at all loci, and a rapid decrease in variance when three co-dominant alleles (at frequencies of $\frac{1}{3}$) are used.

The relationship between the predicted variance, given by inverting the information matrix, and the total variance, estimated by simulation, can be studied by comparing fig. 3 with fig. 5. A summary is provided in table 1, where confidence intervals were excluded for brevity. With more loci, the predicted variance closely matched the total variance, and the fit was particularly good at intermediate gene frequencies and 5 loci. Generally, with fewer loci, extreme gene frequencies or dominance (when

TABLE 1

*Discrepancy between predicted variance per observation (given by formula) and observed variance per observation (based on simulation)*

| | | One locus | | Three loci | | Five loci | |
|---|---|---|---|---|---|---|---|
| | | pred. | obs. | pred. | obs. | pred. | obs. |
| $p = 0{\cdot}5$ | $t = 0{\cdot}9$ | 1·880 | 3·280 | 0·490 | 0·970 | 0·254 | 0·288 |
| | $t = 0{\cdot}5$ | 1·125 | 1·640 | 0·437 | 0·546 | 0·319 | 0·354 |
| | $t = 0{\cdot}2$ | 0·432 | 0·542 | 0·209 | 0·220 | 0·174 | 0·177 |
| | $t = 0{\cdot}05$ | 0·102 | 0·432 | 0·056 | 0·054 | 0·050 | 0·050 |
| $p = 0{\cdot}9$ | $t = 0{\cdot}5$ | 2·830 | 4·000 | 1·013 | 2·140 | 0·654 | 0·992 |
| $p = 0{\cdot}5$ | $t = 0{\cdot}5$, dominance | 2·250 | 4·290 | 0·806 | 1·900 | 0·526 | 0·820 |

the variance is larger) the differences between the predicted and total variance were greater in proportion to the variance. Simulation runs with 10 families and 40 plants per family gave a much closer fit between the observed and predicted, indicating that the effects of small sample size are predominant in the discrepancy.

The estimation procedure also gave small but significant decreases in the variance of ovule gene frequency estimates when more loci were used in the estimate, due to an increased accuracy in inferring the maternal parent. For outcrossing rate and gene frequencies of 0·5, the average probability assigned by the estimator to the true maternal parent at each locus increases when more loci are included in the estimate (table 2). Larger

TABLE 2

*Accuracy in inferring maternal parent at a single locus, measured as average probability assigned by the statistic to the true parent* (t = p = 0·5, *100 plants total*)

| No. of loci | Clegg *et al.* model | | Brown *et al.* model | | |
|---|---|---|---|---|---|
| | 10 per family | 5 per family | Dominance 10 per family | 10 per family | 5 per family |
| 1 | 0·972 | 0·860 | 0·906 | 0·982 | 0·909 |
| 2 | 0·973 | 0·863 | 0·908 | 0·986 | 0·912 |
| 3 | 0·979 | 0·879 | 0·911 | 0·990 | 0·922 |
| 4 | 0·982 | 0·889 | 0·917 | 0·990 | 0·924 |
| 5 | 0·984 | 0·898 | 0·916 | 0·992 | 0·928 |

gains are obtained when there are fewer individuals per family, but not if there is dominance. In using the Brown and Allard (1970) model (which chooses only the most likely maternal genotype to use in subsequent calculations vs. the Clegg *et al.* (1978) model, which includes all likely maternal genotypes) to speed up computations with many loci, we found lower variance in estimates relative to the Clegg *et al.* model. With 10 families and 10 plants per family, the reduction was slight, but when 20 families and 5 plants per family were used, a $\frac{1}{3}$ reduction in variance relative to the model of Clegg *et al.* (1978) occurred. This is due to the merit of choosing only the most likely state, resulting in a higher power in inferring maternal genotypes, demonstrated in the right hand columns of table 2: the average probability assigned to the true parent is greater when the most likely parent is chosen. Thus, it appears desirable to use the most likely maternal parent to increase both computational and statistical efficiency.

### (iii) *Relationship of* n*-locus estimates to lower order estimates*

Since the regularity conditions for asymptotic efficiency of the maximum likelihood estimator as described in this paper are met (cf. Kendall and Stuart, 1979, p. 46), the $n$-locus estimate is the unique minimum variance estimate of $t$ and $\boldsymbol{p}$ using $n$-locus data in the class of all estimators of $t$ and $\boldsymbol{p}$. In other words, any estimate of $t$ and $\boldsymbol{p}$ using subsets of $n$-loci will have a higher variance than the $n$-locus estimate.

Estimates using different combinations of loci but the same plants will be correlated due to the effects of sample outcrossing rate or gene frequency. If we denote $V_a$ as the variance of an estimate $\hat{e}_a$ of $t$ or $p$ based on a subset of

the $n$ loci, $V_b$ as the variance of an estimate $\hat{e}_b$ of the same parameter from the same sample based on another subset of the $n$ loci, and $V_n$ the variance of the $n$-locus estimate of the parameter, the correlation between $\hat{e}_a$ and $\hat{e}_b$ is within the range

$$\rho = \sqrt{\frac{V_n^2}{V_a V_b}} \pm \sqrt{1 - \frac{V_n}{V_a}} \sqrt{1 - \frac{V_n}{V_b}}$$

(Kendall and Stuart 1979, p. 18). If we let one subset be the set of $n$ loci, the correlation between an estimate with variance $V_a$ based on any subset of loci and the $n$-locus estimate with variance $V_n$ is seen to be exactly $\rho = \sqrt{V_n / V_a}$.

## 6. An example

We will compare the estimates of outcrossing in two populations of the annual plant, *Limnanthes alba*: one located 6·7 miles west of Ingot on Hwy 299, Shasta County, California, and the other located 3·1 miles west of Camp Mather in Tuolumne Co., California. Samples of both populations were collected (398 individuals in 36 families from the Ingot population, and 355 individuals in 37 families collected from the Mather population 325) and assayed electrophoretically at several loci, some with 3 alleles (unpublished data of C. I. McNeill). Loci used in this example from the Ingot population are: peroxidase (*Prx*), glutaminoxaloacetic transaminase (*Got* 3) and esterase (*Est* 2); and from the Mather population are: shikimic dehydrogenase (*Sdh*), two Got loci (*Got* 2, *Got* 3) and esterase (*Est* 2).

Data were arranged by families and each individual $n$-locus genotype coded by a string of integers of length $n$, with genotypic assignments viz. for a triallelic locus $A_1A_1 = 1$, $A_1A_2 = 2$, $A_2A_2 = 3$, $A_1A_3 = 4$, $A_2A_3 = 5$, $A_3A_3 = 6$. For example, a 4-locus genotype could be coded as 2153. Loci with dominance (not used here) are coded as $A_1\_ = 1$, $A_2A_2 = 2$, etc. Missing data at a particular locus were coded by a zero; in these cases the estimation procedure is based on a reduced number of loci for that individual only. Separate programs were run to find the variances of estimates by 1) inverting the information matrix and 2) Monte Carlo simulation; these did not differ significantly here due to the large sample size.

Table 3 presents a summary of the estimates using both single locus and multilocus analyses. Single locus estimates varied widely, with some loci indicating a significant between-population difference in outcrossing rate, while the others did not; in addition, $\hat{t}$ based on the *Sdh* locus in the Mather population was significantly lower than the other single locus estimates based on the same sample. The confidence interval was halved by the multilocus estimate and provided a clear distinction in the outcrossing rate between populations. The minimum confidence interval possible (based on the case of infinite loci) is ±0·046 for the Ingot population and 0·031 for the Mather population; these were approached fairly closely by the three locus estimates. The multilocus estimates were slightly higher than the respective means of single locus estimates for both populations indicating that a negative bias has probably been removed by the multilocus estimate. However, even with the increased efficiency of the multilocus estimate, no differences between pollen and ovule frequencies were found.

TABLE 3

*Single and multilocus estimates of outcrossing rate and gene frequencies (with 95 per cent confidence intervals for outcrossing rate[1]), correlations of single locus estimates with the n-locus estimate, and goodness-of-fit statistics[2]*

| Population | Locus | No. sampled | Pollen gene frequencies $\hat{P}_1$ | $\hat{P}_2$ | $\hat{P}_3$ | Ovule gene frequencies $\hat{O}_1$ | $\hat{O}_2$ | $\hat{O}_3$ | Outcrossing rate $\hat{t}$ C.I. | Correlation $r$ | Observed log-likelihood | Expected log-likelihood[3] | No. of times observed exceeded expected[3] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ingot | *Prx* | 392 | 0·65 | 0·35 | — | 0·61 | 0·39 | — | 0·78 ± 0·14 | 0·36 | −335·2 | −326·7 | 61 |
| Ingot | *Got* 3 | 367 | 0·80 | 0·20 | 0·005 | 0·79 | 0·21 | 0·0 | 0·54 ± 0·16 | 0·28 | −203·4 | −214·8 | 175 |
| Ingot | *Est* | 392 | 0·70 | 0·30 | 0·004 | 0·47 | 0·53 | 0·0 | 0·57 ± 0·13 | 0·41 | −314·5 | −306·6 | 78 |
| Ingot | *Prx* | 392 | 0·66 | 0·34 | — | 0·61 | 0·39 | — | | | | | |
| (3 loci) | *Got* 3 | 367 | 0·84 | 0·14 | 0·020 | 0·76 | 0·24 | 0·0 | 0·68 ± 0·08 | — | −861·6 | −791·5 | 6 |
| | *Est* | 392 | 0·63 | 0·37 | 0·002 | 0·47 | 0·53 | 0·0 | | | | | |
| Mather | *Sdh* | 297 | 0·83 | 0·17 | — | 0·72 | 0·28 | — | 0·61 ± 0·16 | — | −242·5 | −173·0 | 0 |
| Mather | *Prx* | 344 | 0·74 | 0·12 | 0·14 | 0·74 | 0·14 | 0·12 | 0·78 ± 0·15 | 0·21 | −318·5 | −315·6 | 117 |
| Mather | *Got* 2 | 335 | 0·67 | 0·33 | — | 0·78 | 0·22 | — | 0·91 ± 0·15 | 0·22 | −266·3 | −244·2 | 11 |
| Mather | *Got* 3 | 328 | 0·69 | 0·31 | — | 0·64 | 0·36 | — | 0·96 ± 0·16 | 0·21 | −269·2 | −255·8 | 38 |
| Mather | *Prx* | 344 | 0·75 | 0·11 | 0·14 | 0·74 | 0·14 | 0·12 | | | | | |
| (3 loci) | *Got* 2 | 335 | 0·72 | 0·28 | — | 0·74 | 0·26 | — | 0·91 ± 0·07 | — | −871·6 | −802·8 | 3 |
| | *Got* 3 | 328 | 0·72 | 0·28 | — | 0·62 | 0·38 | — | | | | | |

[1] Pollen gene frequencies did not differ significantly from ovule frequencies for any of the loci examined here.
[2] Explained in text.
[3] Based on 250 simulated estimates.

To evaluate the goodness of fit of each data set to the multilocus model, Monte Carlo simulation, using the estimates as input parameters, generated log-likelihood values based on 250 data sets. The number of times the log likelihood of the actual estimate exceeded the log likelihood of the simulated estimate formed a basis for acceptance of the model. The single locus estimates in the Ingot population are valid (the last column of table 3 gives the number of simulated log-likelihood values falling below the actual log-likelihood values based on 250 data sets. The number of times the log-likelihood of the actual estimate exceeded the log-likelihood of the simu-simulated log-likelihood clearly exceeds actual value) yielding an aberrant estimate of outcrossing rate at this locus. The 3 locus estimates for both populations are on the borderline of acceptance (*Sdh* was excluded due to its lack of fit, but the 4 locus estimate including *Sdh* was found to be nearly identical to the three locus estimate). The poorer fit of the multilocus data to the model (simulated values fell below actual values only a few times out of 250) relative to the single locus fit, might indicate some linkage between loci, not surprising since this species has a haploid chromosome number of five. As we have demonstrated, the multilocus estimate is less affected by selection and non-random outcrossing than single locus estimates, and one expects this to also hold for linkage, but this goodness of fit test will reject multilocus data even if only one of many loci violates model assumptions.

## 7. Conclusion

We have provided a description of a multilocus model for estimating the outcrossing rate and gene frequencies in plant populations utilizing the family structure of a population. The use of matrix notation has allowed a systematic development of the single locus maximum likelihood methods into a general multilocus method. The basic properties of the multilocus estimate in terms of variance, bias and robustness, using mathematical analysis and computer simulation, indicate the usefulness of this method for determining the proportion of progeny derived from outcrossing. However, the procedure presented here applies to unlinked loci, and its use is not advised for linked loci.

In conclusion, we raise three issues. First, are the improved estimates of outcrossing rates worth the effort required in gathering data at many loci? In relation to allozyme variation studies, multilocus data are often readily available and fewer and slightly smaller families would be adequate as well as a timesaver. Reference to fig. 5 suggests, for example, that three or four loci with intermediate gene frequencies, scored from 200 individuals, will give good estimates. Second, are decreased bias and variance meaningful, when in reality, outcrossing rates vary widely among populations and seasons (*e.g.*, Allard and Workman, 1963; Harding *et al.*, 1974), possibly even among inflorescences within the same plant? Better estimates, in fact, provide a greater statistical rigour in testing such variation in nature. To be sure, differences in outcrossing rates are of primary interest in relation to such matters as the optimal genetic systems, the evolution of inbreeding, and the role of heterosis in inbreeding populations (*e.g.*, the heterozygosity paradox cf. Brown, 1979). Third, do individual loci have different effective outcrossing rates and does the multilocus estimate mask this variability?

The model presented here does not allow interlocus variation in the "effective outcrossing rate" (a parameter summarizing the joint effects of non-random mating, Wahlund effects, and selection upon the production of heterozygosity in the progeny, cf. Allard and Workman, 1963); rather it provides an estimate of the proportion of zygotes derived from cross-fertilization to another genetic individual, even if the mating is between related individuals. The multilocus model presented here can be modified to allow differences between loci in effective outcrossing rates (due to the many degrees of freedom), and should allow a theoretical perspective on the role of variability in effective outcrossing rates among loci in influencing the genetic structure of inbreeding populations.

*Acknowledgments.*—This research was supported in part by grant DEB 7823522 from the National Science Foundation (to S.K.J.) and by a National Research Service Award 5-T32-GMO7467 (NIH) to the senior author. We thank Tony Brown for a critical and helpful review of an early draft and Doug Shaw for providing a copy of his unpublished manuscript.

## 8. References

ALLARD, R. W., AND WORKMAN, P. L. 1963. Population studies in predominantly self-pollinated species. IV. Seasonal fluctuations in estimated values of genetic parameters in lima bean populations. *Evolution, 17,* 470-480.

ALLARD, R. W., KAHLER, A. L., AND WEIR, B. S. 1972. The effect of selection on esterase allozymes in a barley population. *Genetics, 79,* 115-126.

BENNETT, J. H., AND BINET, F. E. 1956. Association between Mendelian factors with mixed selfing and random mating. *Heredity, 10,* 51-55.

BROWN, A. H. D. 1975. Efficient experimental designs for the estimation of genetic parameters in plant populations. *Biometrics, 31,* 145-160.

BROWN, A. H. D. 1979. Enzyme polymorphism in plant populations. *Theor. Pop. Biol., 15,* 1-42.

BROWN, A. H. D., AND ALLARD, R. W. 1970. Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. *Genetics, 66,* 133-145.

BROWN, A. H. D., ZOHARY, D., AND NEVO, E. 1978. Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch. in Israel. *Heredity, 41,* 49-62.

CLEGG, M. T., KAHLER, A. L., AND ALLARD, R. W. 1978. Estimation of life cycle components of selection in an experimental plant population. *Genetics, 89, 765-792.*

FYFE, J. L. AND BAILEY, N. T. J. 1951. Plant breeding studies in leguminous forage crops. I. Natural crossbreeding in winter beans. *J. Agric. Sci., 41,* 371-378.

HARDING, J., MANKINEN, C. B., AND ELLIOTT, M. H. 1974. Genetics of *Lupinus.* VII. Outcrossing, autofertility, and variability in natural populations of the *nanus* group. *Taxon, 23,* 729-738.

JAIN, S. K. 1979. Estimation of outcrossing rates: some alternative procedures. *Crop. Sci., 19,* 23-26.

JAIN, S. K., AND MARSHALL, D. R. 1967. Genetic changes in a barley population analyzed in terms of some life cycle components of selection. *Genetica, 38,* 355-374.

KARLIN, S., AND LIBERMAN, U. 1979. Central equilibria in multilocus systems. I. Generalized nonepistatic selection regimes. *Genetics, 91,* 777-798.

KENDALL, M., AND STUART, S. 1979. *The Advanced Theory of Statistics. Volume 2: Inference and Relationship.* Charles Griffin and Co. Ltd., London.

ROUX, C. Z. 1974. Hardy-Weinburg equilibria in random mating populations. *Theor. Pop. Bio., 5,* 393-416.

SHAW, D. V., AND ALLARD, R. W. 1979. Analyses of mating system parameters and population structure in Douglas fir using single and multilocus methods. In *Isozymes of Forest Trees and Forest Insects.*

SHAW, D. V., KAHLER, A. L., AND ALLARD, R. W. 1980. A multilocus estimator of mating system parameters in plant populations. *Proc. Nat. Acad. Sci. USA, 78,* 1298-1302.

VASEK, F. C. 1968. Outcrossing in natural pouplations: A comparison of outcrossing estimation methods. In: T. Drake (Ed.) *Evolution and Environment.* Yale University Press, New Haven, Conn.

WEIR, B. S., AND COCKERHAM, C. C. 1973. Mixed selfing and random mating at two loci. *Genet. Res., 21,* 247-262.