Theor. Appl. Genet. 61, 321–325 (1982)

Optimum Number of Marker Loci for Estimating Outcrossing in Plant Populations

D. V. Shaw
Department of Genetics, University of California, Davis, Calif. (USA)

A. H. D. Brown
CSIRO, Division of Plant Industry, Canberra, A.C.T. (Australia)

Summary. For the measurements of outcrossing rates in plant populations, current electrophoretic procedures permit many loci to be scored per individual progeny. Given that the total experimental effort or cost is limited, the choice exists then between assaying a large number of loci on a restricted number of individuals, or assaying a large number of individuals at a few loci. Using simple models and the criterion of minimising the variance of the estimate, several factors which affect this choice are considered (levels of polymorphism, heterozygosity, linkage disequilibrium, pollen or outcrossing heterogeneity). The general conclusion is that the actual level of outcrossing is a major factor in determining experimental strategy. Maximum efficiency for estimating outcrossing in predominantly inbreeding plants comes from large samples assayed for few polymorphic loci. In contrast, in predominantly outcrossing plants, more loci should be assayed at the expense of sample size for improved statistical efficiency.

Key Words: Multilocus procedures – Electrophoresis – Breeding systems – Plant population genetics – Statistical efficiency

Introduction

The development of gel electrophoresis of enzymes has provided an ideal technique for the quantitative estimation of outcrossing rates in plant populations (Brown and Allard 1970). Indeed current procedures enable individuals to be scored at an arbitrarily large number of enzyme loci. This has led to the formulation of a number of statistical procedures for estimating outcrossing on a multi-locus basis (Brown et al. 1978; Green et al. 1980; Shaw et al. 1981; Ritland and Jain 1981).

When the experiment is limited by the time and resources available, rather than the amount of plant material, the experimenter faces a basic choice. He can assay a large number of individuals at a few loci, or assay a restricted number of individuals but at many loci. The former procedure will increase the precision of the estimate of outcrossing in a population through scoring more zygotes, whereas the latter increases the precision by increasing the probability of genetically detecting each outcross event. In this paper we examine theoretically how the total experimental effort may be partitioned between the number of plants assayed and the number of marker loci per plant, to estimate outcrossing with maximal precision.

Theoretical Models

The simplest model to begin with assumes a population of a large number of lines homozygous at the marker loci, which are in linkage equilibrium. Suppose an estimate of outcrossing rate (t) is required. We further assume that there is no male gametophytic selection, and that each genotype contributes pollen in frequencies proportional to the maternal genotypic frequencies. From each of m maternal plants, n progeny are tested at k marker loci. The total number of zygotes assayed is N (=nm). If the total experimental effort E (=Nk) is fixed, the optimal value of k is defined as that value which minimises the variance of t for the fixed value of E (Morris and Spieth 1978).

Let F_i denote the population frequency of the i^{th} k-locus maternal homozygous genotype, and G_i ($=1-F_i$, in this model) denote the probability that an outcross will be detected on such a genotype. For this i^{th} maternal genotype, the expected proportion of detectable outcrosses among its progeny is $t\,G_i$, and that of selfs and undetected outcrosses is $1-t\,G_i$. Let x_i denote the total number of observed outcrossed seeds detected on all maternal plants of genotype i, and $(NF_i - x_i)$ denote the selfs or undetected outcrosses. The total number of observed outcrosses is $X=\sum_i x_i$. Finally the $[F_i]$

0040-5752/82/0061/0321/$ 1.00

and therefore the $[G_i]$ are assumed to be known and constant.

The maximum likelihood estimate of t is defined by the solution of the equation

$$\sum_i (NF_i - x_i)\, G_i/(1 - G_i\, t) - X/t = 0 \qquad (1)$$

and the variance of this estimate is approximately given by

$$[\mathrm{var}\,(t)]^{-1} = \sum_i (NF_i - x_i)\,[G_i/(1 - G_i\, t)]^2 + X/t^2 \qquad (2)$$

(from Green et al. (1980) with slight notational changes). If the observed quantities $[x_i]$ in (2) are replaced by their expected values, this formula becomes

$$[\mathrm{var}\,(t)]^{-1} = N \sum_i F_i G_i/t\,(1 - G_i\, t) \qquad (3)$$

More strictly, this formula is the lower bound of the variance (Elandt-Johnson, 1971). The effect of scoring an increasing number of loci (k) is included by substituting $N = Ek^{-1}$, $F_i = (\bar{f}_i)^k$, and $G_i = 1 - F_i = 1 - (\bar{f}_i)^k$ where $\bar{f}_i$ is the geometric mean of the single-locus frequencies for each of the alleles making up the i^{th} k-locus genotype. For example, if $k = 3$ and we consider the 3-locus genotype $A_3B_1C_2$ where A_3 is the third of many alleles at the A locus, etc., then $\bar{f}_{312}$ would be geometric mean of the single-locus allele frequencies of A_3, B_1 and C_2 in the population. Note that $\sum_i (\bar{f}_i)^k = 1$.

$$[\mathrm{var}\,(t)]^{-1} = Ek^{-1} \sum_i (\bar{f}_i)^k\,[1 - (\bar{f}_i)^k]/t\,(1 - t\,[1 - (\bar{f}_i)^k]) \qquad (4)$$

The problem is to define k in terms of t and $(\bar{f}_i)$ such that the R.H.S. of (4) is a maximum. Unfortunately formula (4) is not generally amenable to analysis because the domain of i increases with increasing k.

One approach is to consider simplified examples.

(i) Suppose all loci have 1/p alleles, each with frequency p. This kind of allelic frequency distribution is termed completely "even". The value of $[\mathrm{var}\,(t)]^{-1}$ for $k = 1$ is

$$= E\,(1 - p)/t\,[1 - t\,(1 - p)]$$

which increases as p decreases, and which exceeds the value for $k = 2$, namely

$$E\,(1 - p^2)/2t\,[1 - t\,(1 - p^2)]$$

when $t < (1 + p)^{-1}$. In such cases, $k = 1$ is optimal, giving the lowest variance. Thus in such hypothetical cases, when $t < 0.5$, the optimum value of k is 1 irrespective of p. Table 1 gives optimal values of k for various values of p and t. These values show that the scoring of more equally polymorphic loci per plant is generally efficient only as outcrossing predominates.

Consider the case of one locus with 1/p equally frequent alleles, and a second locus with 1/q equally frequent alleles. The value of $[\mathrm{var}\,(t)]^{-1}$ is:

$$E\,(1 - pq)/2t\,[1 - t\,(1 - pq)]$$

Table 1. Optimum number of equally polymorphic loci at which the alleles are equally frequent (p) for different levels of outcrossing

p	Effective no. of alleles	Outcrossing (t)					
		0–0.6	0.7	0.8	0.9	0.95	0.99
0.50	2	1	2	3	5	6	9
0.33	3	1	1	2	3	4	6
0.25	4	1	1	1 or 2	2	3	5
0.20	5	1	1	1	2	3	4

If the first locus is the more polymorphic of the two (i.e. $p < q$), the question becomes for what levels of outcrossing is it more efficient to score both loci on half the number of plants, rather than scoring only the more variable locus on all the plants. A two-locus strategy is more efficient when:

$$t > (1 - 2p + pq)/(1 - p)\,(1 - pq).$$

Values of outcrossing (t) for various of p and q are shown in Table 2. They confirm the conclusion from Table 1, namely that once the most polymorphic marker locus is known it is theoretically more efficient to score more plants on this locus, than fewer plants on more loci; except in highly outcrossing populations. This conclusion is so far shown for completely even allele frequency distributions.

(ii) Suppose we are dealing with a predominant inbreeder and t can be assumed to be small, t^2 negligible

$$[\mathrm{var}\,(t)]^{-1} \simeq E\,[1 - \sum_i (\bar{f}_i)^{2k}]/tk \qquad (5)$$

The quantity $[1 - \sum_i \bar{f}_i^{2k}]$ is the multilocus diversity parameter (H_e, as used by Brown et al., 1978). Considering now the cases of $k = 1$ and $k = 2$, $[\mathrm{var}\,(t)]^{-1}$ for $k = 1$ exceeds its value for $k = 2$ when

$$2\,[1 - \sum_j p_j^2] > [1 - \sum_j \sum_l p_j^2 q_l^2] \qquad (6)$$

Table 2. Levels of outcrossing (t) for various values of p and q which must be exceeded for a two-locus strategy to be more efficient than the strategy using only single more polymorphic locus (p); assuming all alleles at the polymorphic locus have frequency p, and all alleles at the other locus (with equal or less polymorphism) have frequency q

p	q			
	0.50	0.33	0.25	0.20
0.50	0.67	–	–	–
0.33	0.90	0.75	–	–
0.25	0.95	0.85	0.80	–
0.20	0.97	0.89	0.85	0.83

where $[p_j]$ are allele frequencies at the first locus and $[q_l]$ allele frequencies at the second. This inequality means that if the diversity at the first locus is $1 - \sum_j p_j^2$, there is no gain in precision in scoring half the total number of organism but at this locus as well as at a second locus unless the diversity at the second locus $(1 - \sum_l q_l^2)$ is greater than $(1 - \sum_j p_j^2)/\sum_j p_j^2$. However this would imply that the diversity at the second locus would exceed that of the first. Therefore the most statistically efficient procedure would be to score all organisms at only this more diverse second locus. Another conclusion from this model is that it is desirable to make an initial screening of possible marker loci to learn which has highest diversity. Then the assay for outcrossing should be done using the most variable locus, measured by the diversity parameter.

(iii) Suppose we are dealing with an outcrossing species, with t approaching unity. Then (5) becomes

$$[\mathrm{var}\,(t)]^{-1} \simeq Ek^{-1} \sum_i [1 - (\bar{f}_i)^k] \quad (t \simeq 1.0)$$

$$\simeq Ek^{-1}\,[a_k - 1] \qquad (7)$$

where a_k is the number of k-locus genotypes. This quantity monotonically increases with increasing k. Therefore in such species the most efficient strategy is to assay all the convenient polymorphic marker loci on each seedling, and assay as many seedlings as feasible for these loci.

The Effect of Departure From the Model Assumptions

Several simplifying assumptions were made to construct the above model formula (3). The more important of these are (i) the original population was fully homozygous, (ii) the marker loci were in linkage equilibrium, (iii) there was no sampling variance in the pollen, and (iv) outcrossing was uniform over loci, and (v) the experimental effort required to assay one plant at two loci is the same as that required for two plants at one locus. We consider the effect of relaxing these assumptions on the conclusions from the preceding analysis.

(i) Heterozygosity at a marker locus in any maternal plant markedly reduces the contribution of statistical information made by that locus to the estimate of outcrossing (Jain 1961). Such loci are much less likely to be diagnostic of outcross events. Therefore to the extent that heterozygosity occurs in the maternal plants, the relative efficiency of using more marker loci is increased. In terms of the model presented, the transition from (3) to (4) will not be as simple when heterozygotes are present. The multilocus maternal genotypic frequencies (F_i) can still be assumed as known but the probability of detecting outcrosses on each, G_i, will no longer equal $1 - F_i$, i.e. not all crosses between individuals with different multilocus genotypes will be detectable. The G_i must compensate for the reduced detection capacity associated with heterozygous genotypes. For a given maternal genotype, the expected detection probability, adjusted for the effects of heterozygous loci, will be one minus the product (over loci) of the sums of the frequencies of all alleles present at each locus (Shaw et al. 1981). For example, with three marker loci (A, B, C) each with two alleles (A_1, A_2; B_1, B_2; C_1, C_2) in frequencies (p, q; u, v; l, m) the expected G_i for a maternal individual with the genotype $A_1\,A_1\,B_1\,B_2\,C_2\,C_2$ is:

$$G_i = 1 - (p)\,(u+v)\,(m) = 1 - (p)\,(m).$$

Using this method for calculating G_i and equation 3, the minimum t necessary to justify the use of two equally diverse loci (rather than one locus with twice the number of observations) when heterozygotes are present in the proportions predicted by random mating was obtained (Table 3). As the level of polymorphism decreases the critical value of t asymptotically approaches 0.5. Thus the statement that a multilocus approach is not justified when the outcrossing rate is less than 0.5 seems supported even when large amounts of heterozygosity are included. The effect of multiple alleles (more than two) when heterozygotes are present can also be seen in Table 2. As is the case for completely homozygous populations, detection probabilities are higher for more polymorphic loci (section ii), and the inclusion of additional equally polymorphic loci is justifiable only for higher outcrossing rates.

The analogy can be drawn between lower effective detection probabilities due to heterozygosity, and lower detection probabilities due to lower polymorphism at a locus. The latter effect is seen in Table 1 where the optimum number of loci is higher for a given value of t when the degree of polymorphism is lower (i.e. p higher). In dealing with populations near their inbreeding equilibrium, heterozygosity is higher with higher outcrossing and hence the relative efficiency of multilocus procedures in outcrossed species is reinforced.

(ii) Linkage disequilibrium between the marker loci increases the complexity of the variance of estimates of outcrossing. Brown et al. (1978) previously noted that intense correlation among loci leads to redundant information on oucrossing such that the variance of an estimate based on two loci is higher than if the loci were independent. On the other hand, certain types of "negative" association among alleles may improve the detectability of outcrosses and hence lower the variance more than expected.

To examine these effects in more detail, we consider two marker loci (A, B), each with two alleles (A_1, A_2; B_1, B_2) where their frequencies are (p, q; u, v;

Table 3. The critical value of t which must be exceeded in order to justify the use of two equally diverse loci (both with allele frequencies p, q, and r), at the expense of sample size: heterozygotes are assumed to be present in the proportions predicted by random mating

			Minimum t
p=0.33,	q=0.33,	r=0.34	0.65
p=0.50,	q=0.50,	r=0	0.57
p=0.70,	q=0.30,	r=0	0.55
p=0.90,	q=0.10,	r=0	0.52
p=0.99,	q=0.01,	r=0	0.50

$p < u < v < q$). Maternal plants are fully homozygous, and the frequency of $A_1 A_1 B_1 B_1$ is $pu + D$ where D is the usual linkage disequilibrium parameter ($-pu < D < pv$). Consider the quantity Q (from 3 above):

$$Q = \sum_i F_i G_i / (1 - G_i t) = t\,[N \text{ var } (t)]^{-1};$$

and its behaviour with various values of p, u, D and t. If outcrossing (t) is close to zero, the value of D (symbolised D_m) which maximises Q (i.e. minimises the variance per plant) is:

$$D_m = -(p-q)(u-v)/4 \quad p+u > 0.5$$
$$= -pu \quad p+u < 0.5$$

As t approaches 1.0 closely ($t > 0.99$), Q tends to independence of D except in so far as D affects the number of kinds of maternal genotypes present (formula (7) above), i.e. at the two extreme values of D.

The values of Q for intermediate rates of outcrossing and various gene frequencies were computed. Table 4 summarises the results comparatively. For each set of values of t, p and u, three figures are shown:

(a) D_m, is the value of D which maximises Q. These values are at or close to zero.

(b) The ratio of Q for $D = D_m$ to Q for $D = 0$ indicates the greatest gain in efficiency of estimation when D varies, relative to linkage equilibrium. Since these values are close to 1.0, linkage disequilibrium can only marginally increase efficiency relative to independence.

(c) The ratio of Q for $D = +pv$ to Q for $D = 0$ is the maximum loss in efficiency from altering D. These values are highest in outcrossing species, but indicate that losses can be appreciable in predominant selfers.

It follows that in most situations disequilibrium will generally lower the expected increase in accuracy of estimating outcrossing when increasing number of loci are scored. For inbreeding plant species where disequilibria are common (Brown 1979), this means that few loci and more plants is the better strategy as concluded above.

(iii) Sampling variance in the pollen is assumed to be absent in most multilocus procedures. When pollen allele frequencies depart from those in the maternal plants, either because of selection or sampling effects, estimation becomes more complex. The allele frequencies in the outcrossing pollen must then be estimated. This is likely to be a more severe problem when outcrossing is low, because, for a given number of zygotes assayed, fewer outcrossing pollen grains are included in the sample (Brown, 1979). This problem again adds to the argument against using more loci at the expense of sample size in inbreeding species.

(iv) Differences in estimates of outcrossing derived from different marker loci have been a common problem (Harding and Tucker 1964) especially in the case of morphological markers. For such markers, the differences can be due to a direct effect of morphology outcrossing on rates (Horovitz and Harding 1972). However for isozymes the problem is more likely to stem from sampling and other indirect effects. Heterogeneity of single-locus estimates in outbreeders has therefore been a good reason for using multilocus pro-

Table 4. The effect of linkage disequilibrium on the variance per plant for an estimate of outcrossing (t) based on two loci each with two alleles

Allele frequencies		Levels of outcrossing											
		t=0.01			t=0.5			t=0.9			t=0.99		
p	u	D_m	Gain	Loss	D_m	Gain	Loss	D_m	Gain	Loss	D_m	Gain	Loss
0.1	0.1	−0.01	1.04	0.55	−0.01	1.03	0.33	0.01	1.09	0.52	0.03	1.09	0.41
0.1	0.3	−0.03	1.03	0.88	−0.03	1.02	0.87	0.01	1.01	0.86	0.02	1.02	0.72
0.1	0.5	0	1.0	0.98	0	1.0	0.98	0	1.0	0.89	0	1.0	0.71
0.2	0.4	−0.03	1.01	0.86	−0.01	1.00	0.84	0.01	1.00	0.77	0.02	1.00	0.70
0.3	0.3	−0.04	1.01	0.63	−0.02	1.00	0.55	0.02	1.01	0.41	0.03	1.00	0.34
0.3	0.5	0	1.0	0.87	0	1.0	0.83	0	1.0	0.73	0	1.0	0.68
0.4	0.4	−0.01	1.00	0.66	0.00	1.00	0.56	0.01	1.0	0.40	0.01	1.00	0.34
0.5	0.5	0	1.0	0.67	0	1.0	0.56	0	1.0	0.39	0	1.0	0.34

D_m = value of disequilibrium for minimum variance per plant; Gain = efficiency increase for D_m relative to linkage equilibrium ($D = 0$); Loss = efficiency relative to $D = 0$, when D is such as to maximise the variance per plant

cedures (Shaw et al. 1981), and reinforces our conclusions here concerning optimal procedures. For inbreeders, this problem may indicate that estimates based on a few isozyme markers are more desirable than just a single-locus estimate (e.g., Green et al. 1980), despite their theoretical inefficiency.

(v) The relative effort required to increase the number of loci scored is generally not equivalent to increasing the sample size. Electrophoretic procedures generally enable the assay of several loci on a single gel without much increase in effort. Also the number of seed available may be limited. Further, errors can occur in scoring genotypes from zymograms. In predominant selfers, the rare outcrosses may have to be checked by progeny testing unless it is multiply marked (Brown et al. 1978). Therefore our simplifying assumption does not take into account these benefits of an experimental strategy based on more than the single most variable marker loci.

However the methodology developed can still be used in partitioning effort, provided that information is available as to allele frequencies at the potentially assayable loci. For example if equal effort is expended for each starch gel, the relevant question may be whether to obtain one or two sets of loci for each zygote. This question can be analyzed with equation 3 as modified to include the effects of heterozygosity. The detection probabilities, G_i, for each set of loci can be compared in ways directly analogous to the two locus comparisons in Tables 2 and 3.

Conclusions

Many factors affect the optimum number of marker loci to use when estimating outcrossing in a plant population. The simplified theoretical model above however suggested that the actual level of outcrossing is the major factor. In inbreeders, maximum theoretical efficiency obtains when only the most variable single-locus is used and as many plants as possible are scored for this locus. In contrast, when the level of outcrossing is very high, it is theoretically more efficient to score many loci at the cost of scoring few individuals at all the loci. In more complex situations, this general conclusion appears to be reinforced.

Literature

Brown, A.H.D. (1979): Enzyme polymorphism in plant populations. Theor. Popul. Biol. **15,** 1–42

Brown, A.H.D.; 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.; Nevo, E. (1978): Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch in Israel. Heredity **41,** 49–62

Elandt-Johnson, R.C (1971): Probability Models and Statistical Methods in Genetics. New York: Wiley

Green, A.G.; Brown, A.H.D.; Oram, R.N. (1980): Determination of outcrossing in a breeding population of *Lupinus albus* L. Z. Pflanzenzücht. **84,** 181–191

Jain, S.K. (1961): A note on the estimation of natural crossing by the maximum likelihood method. Ind. J. Genet. Plant Breed. **21,** 146–148

Harding, J.; Tucker, C.L. (1964): Quantitative studies on mating systems. 1. Evidence for the non-randomness of outcrossing in *Phaseolus lunatus* Heredity, **19,** 369–381

Horovitz, A.; Harding, J. (1972): Genetics of *Lupinus* V. Intraspecific variability for reproduction traits in *Lupinus nanus.* Bot. Gaz. **133,** 155–165

Morris, R.W.; Spieth, P.T. (1978): Sampling strategies for using female gametophytes to estimate heterozygosity in conifers. Theor. Appl. Genet. **51,** 217–222

Ritland, K.; Jain, S. (1981): A model for the estimation of outcrossing rate and gene frequencies using independent loci. Heredity, **47,** 35–52

Shaw, D.V.; Kahler, A.L.; Allard, R.W. (1981): A multilocus estimator of mating system parameters in plant populations. Proc. Nat. Acad. Sci. (USA) **78,** 1298–1302

Received November 20, 1981
Communicated by R. W. Allard

Dr. D. V. Shaw
Department of Genetics
University of California
Davis, Calif. 95616 (USA)

Dr. A. H. D. Brown
CSIRO
Division of Plant Industry
P. O. Box 1600
Canberra A.C.T. 2601 (Australia)