Theor Appl Genet (1994) 88:925–932

Y. B. Fu · K. Ritland

On estimating the linkage of marker genes to viability genes controlling inbreeding depression

Received: 28 July 1993 / Accepted: 12 January 1994

Abstract Statistical properties and extensions of Hedrick and Muona's method for mapping viability alleles causing inbreeding depression are discussed in this paper. Their method uses the segregation ratios among selfed progeny of a marker-locus heterozygote to estimate the viability reduction, "*s*", of an allele and its recombination fraction, "*c*", with the marker. Explicit estimators are derived for *c* and *s*, including expressions for their variances. The degree of estimation bias is examined for cases when (1) the viability allele is partially recessive and (2) the marker locus is linked to two viability loci. If linkage or viability reduction is moderate, very large sample sizes are required to obtain reliable estimates of *c* and *s*, in part because these estimates show a statistical correlation close to unity. Power is further reduced because alleles causing viability reduction often occur at low frequency at specific loci in a population. To increase power, we present a statistical model for the joint analysis of several selfed progeny arrays selected at random from a population. Assuming a fixed total number of progeny, we determine the optimal number of progeny arrays versus number of progeny per array under this model. We also examine the increase of information provided by a second, flanking marker. Two flanking markers provide vastly superior estimation properties, reducing sample sizes by approximately 95% from those required by a single marker.

Key words Linkage · Marker genes · Viability genes
Distorted segregation · Inbreeding depression

Introduction

Many studies have demonstrated that genetic markers may show distorted (non-Mendelian) segregation ratios when linked to loci affecting viability (Sorensen 1967; Grant 1975; Zamir and Tadmor 1986; Weeden and Wendel 1990; Lyttle 1991; Wagner et al. 1992). Hedrick and Muona (1990) showed how such distorted segregation ratios can be used to characterize viability alleles in self-fertile organisms. If a plant with a heterozygous marker locus (A_1A_2) linked to a heterozygous viability locus is selfed, one can use the frequencies of the three progeny marker genotypes (A_1A_1, A_1A_2, A_2A_2) to estimate the viability reduction at the linked locus "*s*", as well as the recombination fraction between the marker and viability loci "*c*". Because there are only two degrees of freedom in the data, one must assume that the viability allele is completely recessive, as opposed to being partially recessive. Using their method, they found a near-lethal allele at a locus closely linked to an *Esterase* marker in Scots pine (*Pinus sylvestris* L.).

Their method raises the prospect that the genetic basis of inbreeding depression in self-fertile species can be studied at a resolution not attainable by classical methods (c.f. Savolainen et al. 1992 for summary), yet without the intensive labor involved with constructing saturated linkage maps, as only single loci are needed. With the advent of molecular markers (e.g., Botstein et al. 1980; Williams et al. 1990; Herne et al. 1992) this method shows promise for identifying and characterizing individual loci causing inbreeding depression in plants.

However, the statistical properties of mapping recessive viability loci using single marker loci have not been considered. In general, experimental designs and statistical power have only been investigated for methods used in mapping fecundity loci (e.g., Soller et al. 1976; Lander and Botstein 1989; Knapp et al. 1990; Carbonell et al. 1993; Darvasi et al. 1993; Van der Beek and Van Arendonk 1993). While mapping fecundity loci involves the comparisons of quantitative trait means among marker genotypes, mapping *viability* loci involves analyses of marker genotype frequencies. The use of frequencies, as opposed to means, may result in different estimation properties. In addition, the robustness of estimating recessive viability alleles via single-marker loci warrants a study. For example, in a recent study, a graphical analysis of marker segregation in

Communicated by A. L. Kahler

Y. B. Fu · K. Ritland (✉)
Department of Botany, University of Toronto, Toronto, Ontario M5S 3B2 Canada

31 selfed progeny arrays of *Mimulus guttatus* plants indicated that viability alleles were most often partially dominant (Fu and Ritland 1994). Such dominance may bias the estimation of c and s.

In this paper, we first derive explicit estimators for c and s using Hedrick and Muona's (1990) method and provide expressions for their variances. We then discuss the degree of estimation bias for those cases when the viability allele is partially recessive and when the marker is linked to two selected loci, and give sample sizes required to detect linkage. Furthermore, we explore experimental strategies for increasing the power of estimating the linkage of marker loci to recessive viability loci. Two alternatives are considered: a statistical model for jointly analyzing several selfed progeny arrays and a flanking marker model for characterizing the increased information for the power of detection.

Estimators and their variances

Estimators

Expanding upon the work of Hedrick and Muona (1990), we first derive the estimators of c and s, and formulas for their variances. Let us denote a plant with a heterozygous marker locus linked to a heterozygous viability locus as A_1B_d/A_2B_+, where A is the marker locus, B the selected locus, and d is the completely recessive, selected allele. If this plant is selfed, the expected frequencies of its progeny marker genotypes A_1A_1, A_1A_2, and A_2A_2 are:

$$p_{11} = \frac{1-s+2sc-sc^2}{4-s}$$
$$p_{12} = \frac{2(1-sc+sc^2)}{4-s} \qquad (1)$$
$$p_{22} = \frac{1-sc^2}{4-s},$$

respectively (Hedrick and Muona 1990). Since Eq. 1 is a system of two independent equations with two unknowns, we can obtain the estimators as exact solutions of Eq. 1 as:

$$\hat{c} = \frac{2P_{22}-P_{12}}{2(P_{22}-P_{11})} \qquad (2)$$
$$\hat{s} = \frac{4(P_{22}-P_{11})^2}{(1+P_{22}-P_{11})^2-4P_{22}},$$

where uppercase P's indicate *estimated* frequencies – those observed in the experiment. Since Eq. 1 is quadratic in c, there is a second solution, but this solution involves negative values of c.

Variance of estimates

The variances and covariance of estimates are found by inverting the Fisher information matrix, whose elements consist of expected second derivatives of the log-likelihood function (Stuart and Ord 1991). Using an analytical equation solver, we found these variances and covariance to be

$$V(\hat{c}) = \frac{[sc(4c^3-8c^2+3c+1)-(4c^2-4c+3)](s-4)}{2s^2(2c-1)^2}$$
$$V(\hat{s}) = \frac{[sc^4(s-8)-2sc^3(s-8)+c^2(s^2-12s+8)+4c(s-2)+3-s](s-4)^2}{(2c-1)^4} \qquad (3)$$
$$Cov(\hat{c},\hat{s}) = \frac{[2sc^4-4sc^3+2c^2(s-1)+2c-1](s-4)^2}{s(2c-1)^3}.$$

The denominator of Eq. 3 shows that linkage ($c< 0.5$) is required for estimation and that weak linkage causes high variance. Numerical examination of the correlation coefficient between c and s over various parameter sets reveals that $\hat{c}$ and $\hat{s}$ have a very high statistical correlation, with a coefficient usually greater than 95%. This has unfortunate consequences for the joint estimation of c and s, as discussed below.

LOD scores

These estimates maximize the LOD score

$$LOD = N_{11}\log_{10}\left[\frac{P_{11}}{0.25}\right] + N_{12}\log_{10}\left[\frac{P_{12}}{0.50}\right] + N_{22}\log_{10}\left[\frac{P_{22}}{0.25}\right], \qquad (4)$$

where N_{ij} is the number observed for each marker genotype. Following the common practice in gene mapping, Hedrick and Muona (1990) stated that a LOD value in excess of 3 indicates significant linkage of a viability allele. This test is actually very conservative if applied to a single marker locus because by the likelihood ratio test criterion, $-2\ln(L_1/L_0)=-4.30$ LOD is distributed as χ^2 with 2 degrees of freedom (note that the left side is the natural log and the right side $\log_{10}$). Thus, a LOD value greater than just 1.3 indicates significance at the 95% level. This lower threshold for significance is acceptable if only a single marker locus is examined. If more markers are examined in the same experiment, a higher LOD threshold is required; with a large number of markers, the correct critical value depends upon the total map distance in a genome.

The space of allowed segregation

The pattern of distorted segregation of a marker linked to a viability locus can be precisely given by the range of the *joint* distribution of progeny genotype frequencies (p_{11}, p_{12} , p_{22}) in Eq. 1, wherein the bounds for c and s place bounds upon the p's. Assuming the linkage phase A_1B_d/A_2B_+ as before, the range of p_{12} is

$$\frac{1}{2} \le p_{12} \le \frac{2}{3} \tag{5a}$$

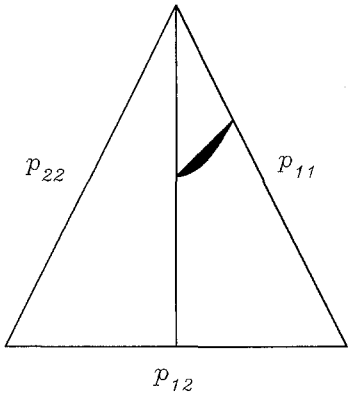
(these are the allowable values of p_{12} in Eq. 1 for $0 \le c \le 0.5$ and $0 \le s \le 1.0$). The allowable range for p_{22}, given p_{12}, is

$$\frac{p_{12}}{2} \le p_{22} \le \frac{1}{2} - \frac{p_{12}}{2} + \frac{\sqrt{6\,p_{12} - 3}}{6} \tag{5b}$$

(this was obtained by examination of limits for $p_{12}+p_{22}$ in Eq. 1). At this point, p_{11} is determined. The ranges for the *marginal* distributions of progeny frequencies are $1/2 \le p_{12} \le 2/3$, $1/4 \le p_{22} \le 1/3$, and $0 \le p_{11} \le 1/4$. Equations 5a and b show that genotypes homozygous for markers coupling with deleterious alleles are always least frequent and heterozygotes always in excess. This is consistent with the findings of Zamir and Tadmor (1986) and Wagner et al. (1992) for the special case of $s=1$.

Figure 1 graphically depicts the range of possible segregation ratios defined by Eqs. 5a and b. The "space" of allowed segregation under the completely recessive model of Hedrick and Muona is indicated in shadow. Regions outside this space (but within the right half of the triangle) are occupied by other selection models, such as partial dominance and overdominance (see Fu and Ritland 1994). Figure 1 shows that the "space" of segregation ratios under the completely recessive model is quite small and oblique. It is likely that the high statistical correlation between the estimates of c and s (Eq. 3) is related to the obliqeness of this allowed space of segregation.

Fig. 1 The space of segregation ratios under the completely recessive selection model of Hedrick and Muona, as shown in *shadow*. The frequency of any one genotype is obtained by drawing a line perpendicular to the axis labeled by that genotype out to the point within the triangle representing the segregation ratio (only the right half of the equilateral triangle is occupied because allele 1 when homozygous is arbitrarily defined to have lower viability than allele 2 when homozygous)



Bias due to partial dominance

Alternatives to the completely recessive model of Hedrick and Muona include selection models of additivity or partial dominance (Simmons and Crow 1977; Charlesworth and Charlesworth 1987; Fu and Ritland 1994). Such alternatives are "violations" of the completely recessive model when the latter is assumed. In these cases, estimates of c and s using Hedrick and Muona's method are biased. To investigate these biases, we formulated a more general selection model with fitness regime $1-s$, $1-hs$, and 1 for genotypes B_sB_s, B_sB_+, and B_+B_+, respectively, at the selected locus, where h is the degree of dominance. In this model, the expected frequencies of marker genotypes A_1A_1, A_1A_2, and A_2A_2 in progeny of a selfed double heterozygous plant A_1B_s/A_2B_+ are

$$p_{11} = \frac{c^2 s(2h-1) - 2cs(h-1) + (1-s)}{4 - s(2h+1)}$$

$$p_{12} = \frac{2(-c^2 s(2h-1) + cs(2h-1) - hs + 1)}{4 - s(2h+1)} \tag{6}$$

$$p_{22} = \frac{c^2 s(2h-1) - 2csh + 1}{4 - s(2h+1)}$$

This was obtained by specifying the two-locus frequencies in selfed progeny, then summing over genotypes at the selected locus B. If $h=0$, the selection model becomes equivalent to the completely recessive model of Hedrick and Muona. Bias is obtained by substituting genotypic frequencies of Eq. 6 into Eq. 2 and subtracting these estimates from their true values. This gives

$$\text{bias}(\hat{c}) = h(1-2c)$$

$$\text{bias}(\hat{s}) = \frac{hs(2-hs)}{1-h(2-hs)}. \tag{7}$$

Equation 7 shows that the bias of both estimates is usually positive. For example, when $h=0.20$, $c=0.10$, and $s=0.65$, the bias of $\hat{c}$ is 0.16 and the bias of $\hat{s}$ is 0.39. Equation 7 also shows that the bias of both estimates increases approximately linearly with h and that the bias of $\hat{c}$ is worst at low values of c while the bias of $\hat{s}$ is greater at high levels of s. Interestingly, the bias of $\hat{s}$ is independent of linkage c.

With a moderate level of dominance $h=0.20$, a deleterious allele of large effect (e.g., $s=0.65$) will tend to be detected as a completely lethal allele ($\hat{s}=1.04$). However, we note that for lethal or sublethal alleles, the level of dominance is probably low ($h<0.03$; Simmons and Crow 1977; Charlesworth and Charlesworth 1987), so that, in general, biases for these alleles are small. For example, when $h=0.02$, $c=0.05$, and $s=0.95$, the bias of $\hat{s}$ is only 0.04.

Bias due to two selected loci

The procedure of Hedrick and Muona (1990) also assumes the presence of just one selected locus in the vicinity of the

marker locus. If one marker is linked to more than one selected locus, estimation bias may occur. In the light of recent findings of the infrequent linkage of marker and viability loci (Fu and Ritland 1994), one might expect that the linkage of multiple viability loci is uncommon, but it would still be of interest to understand the degree of such an estimation bias, if any. For this reason, we examined the bias due to the linkage of two selected loci. For such a linkage, one must specify whether (1) these loci flank the marker or reside on one side of the marker, and (2) the selected alleles are in coupling or repulsion with respect to each other. Because of this complexity, we resorted to Monte-Carlo simulations wherein 5,000 zygotes were randomly generated according to locus arrangement, strength of selection, and recombination fraction. A multiplicative fitness model was assumed, with the fitness of a double homozygote being $(1-s_1)(1-s_2)$. Estimation was applied using Eq. 2, and the entire procedure replicated 1,000 times.

Table 1 gives the resulting biases of $\hat{c}$ and $\hat{s}$ in ten representative cases. We first discuss the case of selected loci that flank the marker locus. When flanking selected loci are in coupling and $c_1=c_2$, estimates of c show positive bias. When $c_1 > c_2$, they tend towards the larger, c_1. The estimated s shows a positive bias when either $s_1=s_2$ or $s_1<s_2$. When flanking selected loci are in repulsion, both $\hat{c}$ and $\hat{s}$ show extreme bias.

The pattern of bias is more complex when the selected loci are non-flanking with respect to the marker. When selected loci are in coupling, both $\hat{c}$ and $\hat{s}$ show positive upward biases in all three situations examined ($c_1=c_2$, $c_1>c_2$, and $c_1<c_2$). Severe bias is found when selected loci are in repulsion.

In summary, the presence of two selected loci in coupling generally results in estimates not representative of either selected locus. Both average c and s are usually overestimated. When the selected loci are in repulsion, extreme biases occur, with apparent overdominance of the neutral marker.

Table 1 The biases caused by two selected loci linked to a marker locus

	True value				Estimated value			
					(Selected loci in coupling)		(Selected loci in repulsion)	
	c_1	s_1	c_2	s_2	$\hat{c}$	$\hat{s}$	$\hat{c}$	$\hat{s}$
Flanking								
	0.20	0.90	0.20	0.90	0.26	1.53	0.52	0.00
	0.10	0.90	0.10	0.90	0.15	1.29	1.97	0.00
	0.05	0.90	0.05	0.90	0.08	1.15	–3.18	0.00
	0.30	0.45	0.05	0.90	0.11	1.08	1.15	0.59
	0.20	0.45	0.05	0.90	0.11	1.09	1.34	0.42
	0.10	0.45	0.05	0.90	0.09	1.06	1.64	0.28
Non-flanking								
	0.20	0.90	0.20	0.90	0.27	1.25	–0.62	0.16
	0.10	0.90	0.10	0.90	0.16	1.13	–2.99	0.04
	0.05	0.90	0.05	0.90	0.09	1.07	–8.16	0.01
	0.05	0.90	0.30	0.45	0.10	1.05	–0.12	0.63
	0.05	0.90	0.20	0.45	0.10	1.05	–0.27	0.48
	0.05	0.90	0.10	0.45	0.09	1.02	–0.50	0.34
	0.30	0.45	0.05	0.90	0.31	1.02	1.20	0.17
	0.20	0.45	0.05	0.90	0.23	1.01	1.51	0.16
	0.10	0.45	0.05	0.90	0.14	0.99	1.81	0.16

Sample sizes required to reject the null hypothesis of no selected locus

Under the completely recessive model, the null hypothesis of no linkage of a marker with a selected locus can be stated in two alternative ways. First, it can be stated in terms of the model parameters, as $c=0.50$ and/or $s=0.00$. The confidence interval of either estimate, obtained from Eq. 3, is tested for overlap with these values. Because estimates of c and s are highly correlated, their tests will be highly correlated, so that the choice of c versus s is rather arbitrary. Second, the null hypothesis can be stated in terms of the marker locus ratio, as the null ratio 1:2:1 and the LOD score (Eq. 4) is used to test for deviation from 1:2:1. This is a more general test that actually tests for the presence of *any* model of selection (including dominance and overdominance).

We first consider the expected sample size required to reject the null hypothesis, $s=0.00$, 95% of the time. We calculated sample sizes required to exclude the confidence interval for $\hat{s}$ from 0, or the value of N satisfying

$$\hat{s} - 1.96\sqrt{\frac{V(\hat{s})}{N}} = 0, \tag{8}$$

where $V(\hat{s})$ is the variance of $\hat{s}$ (Eq. 3). The sample sizes required to reject this null hypothesis are shown in Fig. 2a. Because sample sizes depend upon the true values of both c and s, three representative values of s are plotted against the range of possible c values.

In general, Fig. 2a shows that very large sample sizes are needed for a reasonable statistical power under the Hedrick and Muona's model. For example, an experiment with a sample size of approximately 1,000 can be expected to detect only the close linkage ($c< 0.10$) of loci with $s>0.40$. An experiment with a sample of approximately 100 can be expected to detect only close linkage of near lethal alleles ($s\approx 1$). This accords with the detection by Hedrick and Muona (1990) of a near lethal ($\hat{s}=0.896$) allele closely linked ($\hat{c}=0.084$) to the *Esterase* marker, based upon a sample of only 75 genotypes.

Next, we consider the expected sample size required to reject the null hypothesis of a 1:2:1 ratio at the marker locus 95% of the time. Following Eq. 4, we calculated sample sizes N satisfying

$$N\left(p_{11}\log_{10}\left[\frac{p_{11}}{0.25}\right] + p_{12}\log_{10}\left[\frac{p_{12}}{0.50}\right] + p_{22}\log_{10}\left[\frac{p_{22}}{0.25}\right]\right) = 1.3. \tag{9}$$

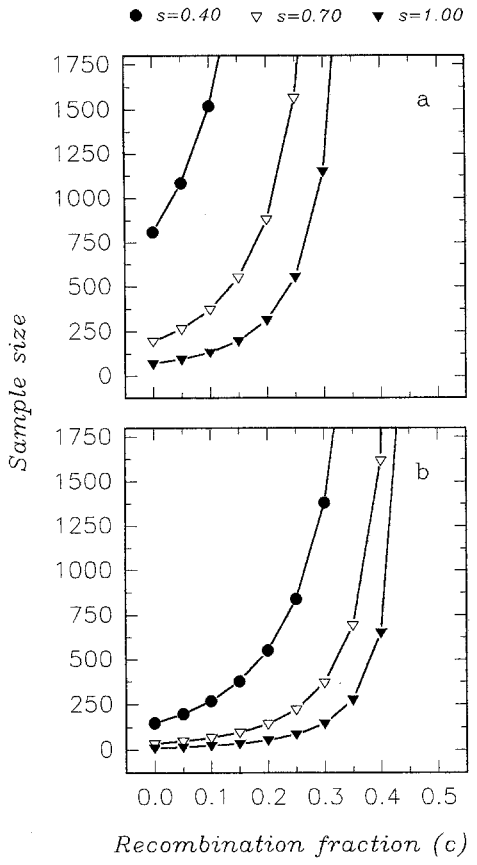


Fig. 2a, b Sample sizes required to reject the null hypothesis of no linkage to a viability allele for three levels of deleterious effects. These sample sizes are determined on the basis of **a** the simultaneous estimates of c and s and **b** the LOD test for deviation of the 1:2:1 ratio

These values of N are shown in Fig. 2b for the same set of parameter values as in Fig. 2a. Figure 2b shows that, in contrast, a much smaller sample size (5%–20% of the number in Fig. 2a) is sufficient to reject the null hypothesis of no distorted segregation. But this is because the alternative model is *any* model of selection, ranging from incomplete dominance to overdominance. Thus, while enticing in its properties, the LOD test is not appropriate unless a general model of selection is invoked.

Increasing power via multiple arrays of selfed progeny

In natural populations, the frequency of a deleterious allele is usually quite low (Charlesworth 1992). Alleles of low frequency are not likely to be present in a sample consisting of a single parent. This facet of statistical power, e.g., the sampling of viability alleles from a population, was not incorporated into the above calculations (which assumed that a viability allele was present in a single, sampled parent). To incorporate the sampling of alleles from a parent population, a model involving several selfed progeny arrays is required.

In this section, we present a statistical procedure, an extension of Hedrick and Muona's model, that incorporates the frequency of the selected allele. This procedure is based upon the assay of several selfed progeny arrays. We assume a single selected allele in the population, with the same effect s among all individuals as well as the same recombination fraction c with the associated marker locus. This selected allele exists in a heterozygous condition at frequency f. The likelihood of the data "D" from several selfed progeny arrays, wherein the k-th array has data D_k consisting of $N_{11,k}$, $N_{12,k}$, $N_{22,k}$ progeny of genotypes A_1A_1, A_1A_2, and A_2A_2, respectively, is the mixture model

$$L(D)=\Sigma_k \ln [(1-f)\, L(D_k \mid \text{hom}) + f\, L(D_k \mid \text{het})], \qquad (10)$$

where the likelihood of the data assuming that the parent is homozygous for this selected locus is

$$L(D_k \mid \text{hom}) = (1/4)^{N_{11,k}} (1/2)^{N_{12,k}} (1/4)^{N_{22,k}}$$

(N_{ij} is the number of each respective marker genotype) and the likelihood of the data assuming that the parent is heterozygous at the selected locus is

$$L(D_k \mid \text{het}) = p_{11}^{N_{11,k}}\, p_{12}^{N_{12,k}}\, p_{22}^{N_{22,k}}$$

(p_{ij} is given by Eq. 1).

For two major reasons, we now discuss with this model the optimal allocation of experimental resources. First, experimental resources usually limit the assay of total number of progeny such that the assay of more progeny arrays reduces the number of progeny per array, and vice-versa. Second, with too few progeny arrays assayed, one may waste effort assaying non-informative arrays (homozygotes at the viability locus), but if too few progeny per array are assayed, alleles of small effect cannot be detected. To investigate an optimal allocation, we performed Monte-Carlo simulations. Data were simulated by randomly choosing parentage (heterozygous with probability $2q(1-q)$, where q is the frequency of the selected allele, otherwise homozygous), and then randomly choosing progeny with probabilities p_{11}, p_{12}, and p_{22} (if heterozygous) or 1/4, 1/2, and 1/4 (if homozygous). Estimates of f, c, and s were then found by numerically maximizing Eq. 10. For each set of parameter values, 10^5 replications were performed. The power of detecting the selected allele was measured as the proportion of estimates of s greater than zero (whichever parameter – f, c, or s – was used, the power showed the same trend).

Figure 3 shows the power or probability of detecting a selected allele, for seven allocations. Total sample size was 2,048, and six combinations of parameters were considered: (f=0.05, 0.20)×(c=0.00, 0.25)×(s=1.00, 0.70). Figure 3a shows that for f=0.05, the optimal allocations vary from 64 progeny arrays/32 progeny per array at c=0.00 and s=1.00 to 16 progeny arrays/128 progeny per array at c=0.25 and s=0.70. Thus, as the allelic effect and/or linkage decrease, we require more progeny per array at the expense of fewer progeny arrays; the power also decreases. Figure 3b shows that when f=0.20, optimal allocations involve fewer progeny arrays for the same parameter values; the overall power also increases. We also examined allo-

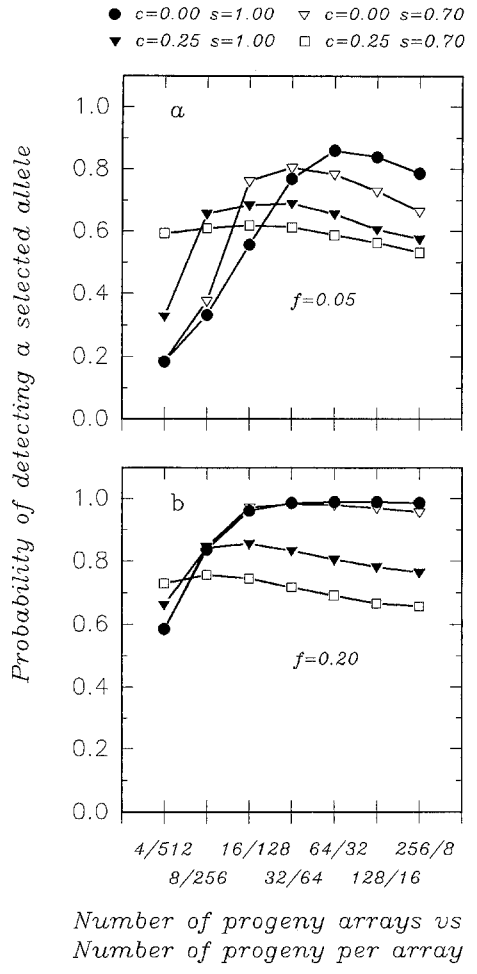


Fig. 3a, b Probabilities of detecting a selected allele for seven combinations of number of selfed progeny arrays versus number of selfed progeny per array in a given sample of 2,048 progeny with six sets of parameter values. These probabilities are shown separately for the heterozygous parent frequencies (f) of **a** 0.05 and **b** 0.20

cations with total sample sizes of 1,024 and 4,096 and found similar optimal allocations for the ratio, (number of progeny arrays)/(number of progeny per array).

These results show that many progeny arrays are required to detect a selected allele with low frequency. While the number of progeny arrays depends upon the frequency, linkage, and selective effect of the allele, all of which are unknown until after the experiment is conducted, as a rule of thumb, we recommend 16–32 selfed progeny arrays in an experiment of 2,000 progeny total. For other sample sizes, the optimum for the ratio (number of progeny arrays)/(number of progeny per array) lies in the range of 1/8 to 1/2.

Increasing power via two flanking marker loci

To increase the power of detection, one could also assay more marker loci in the vicinity of the selected locus. To address this, we computed the power to detect linkage of a selected locus linked to two flanking markers as follows. The two marker loci are denoted A and B, and the selected locus Q. The recombination fraction between Q and A is c_1 and the recombination fraction between Q and B is c_2. If a heterozygous plant of genotype *AQB/aqb* is selfed, there are nine possible marker genotypes; the probabilities before selection are given in Table 2. These are divided into marginal (p_i) and conditional (π_{ij}) probabilities, where i refers to marker genotypes (i=1, ... , 9) and j to viability genotypes (j=1, ... , 3). We assume the Haldane function with no interference, and define $c=c_1+c_2-2c_1c_2$, $r_1=c_1c_2/(1-c)$, and $r_2=c_1(1-c_2)/c$ (c is the recombination fraction between the markers). The expected frequencies of the marker genotypes after recessive selection are $P_i'=P_i/(\Sigma P_i)$, for i=1, ... , 9 and where $P_i=p_i(\pi_{i1}(1-s)+\pi_{i2}+\pi_{i3})$. Overall, there are three parameters to estimate (c_1, c_2, and s), and 8 degrees of freedom in the data.

Table 2 Marginal (p_i) and conditional (π_{ij}) probabilities of the three genotypes of the viability locus for a given flanking marker genotype

Marker genotype (i)	p_i	QQ π_{i1}	Qq π_{i2}	qq π_{i3}
AABB	$(1-c)^2/4$	$(1-r_1)^2$	$2r_1(1-r_1)$	r_1^2
AABb	$c(1-c)/2$	$(1-r_1)(1-r_2)$	$r_1+r_2-2r_1r_2$	r_1r_2
AAbb	$c^2/4$	$(1-r_2)^2$	$2r_2(1-r_2)$	r_2^2
AaBB	$c(1-c)/2$	$r_2(1-r_1)$	$1-r_1-r_2+2r_1r_2$	$r_1(1-r_2)$
AaBb	$(1-c)^2/2+c^2/2$	$r_1(1-r_1)/2+r_2(1-r_2)/2$	$[r_1^2+(1-r_1)^2]/2+[r_2^2+(1-r_2)^2]/2$	$r_2(1-r_2)/2+r_1(1-r_1)/2$
Aabb	$c(1-c)/2$	$r_1(1-r_2)$	$1-r_1-r_2+2r_1r_2$	$r_2(1-r_1)$
aaBB	$c^2/4$	r_2^2	$2r_2(1-r_2)$	$(1-r_2)^2$
aaBb	$c(1-c)/2$	r_1r_2	$r_1+r_2-2r_1r_2$	$(1-r_1)(1-r_2)$
aabb	$(1-c)^2/4$	r_1^2	$2r_1(1-r_1)$	$(1-r_1)^2$

Figure 4 shows the sample sizes required to detect a allele in a heterozygous parent with 95% probability. Figure 4a was based upon applying Eq. 8, using the variance of s obtained by inverting the information matrix for c_1, c_2, and s. Figure 4b was based upon the LOD criteria, $N\Sigma P_i'\log_{10}(P_i'/p_i)>1.3$. Note that the difference in the degree of freedom between the numerator and the denominator in this LOD score is the same as in Eq. 9, namely 2, since under the null hypothesis the two marker loci may still be linked, i.e., $c<0.50$.

The comparison of Fig. 4a to Fig. 2a shows that flanking markers provide much more efficient detection, as the required sample sizes are reduced to 2–5% from those required by a single marker. The pattern of efficiency depends upon the selective effect s. At s=0.40, flanking markers are 16 times more efficient when markers are 10 map units away and 36 times more efficient when markers are 30 map units away. At s=1.00, the opposite trend occurs for map distance; flanking markers are 293 times more efficient when markers are 10 map units away and 37 times more efficient when markers are 30 map units away. At s=0.70, there is almost no dependence upon map distance

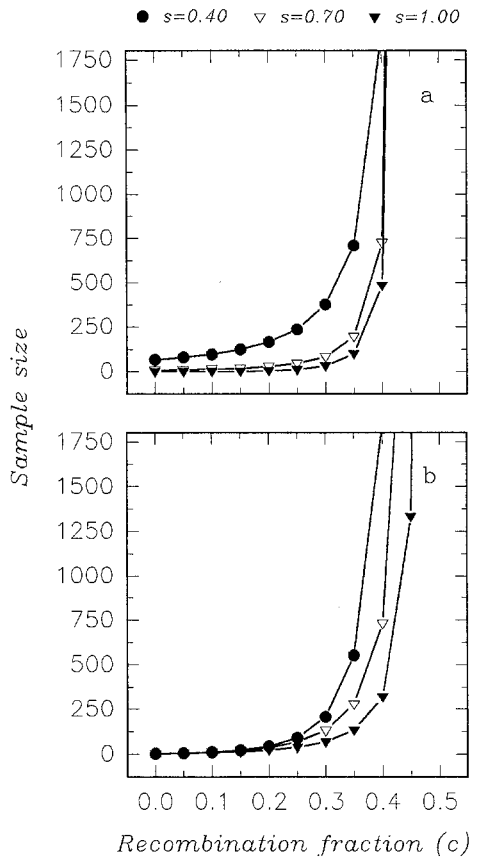


Fig. 4a, b Sample sizes required to reject the null hypothesis of no linkage to a viability allele for three levels of deleterious effects when a second flanking marker is used. These sample sizes are determined, on the basis of **a** the simultaneous estimates of *c* and *s* and **b** the LOD criteria

(30–40 times more efficient). The pattern for the LOD criteria (Fig. 4b) is similar.

Discussion

While the sample sizes required to detect deviations from 1:2:1 are relatively small, the sample sizes needed to reliably detect recessive viability alleles using Hedrick and Muona's method can be quite large (Fig. 2). This large discrepancy occurs because the former is effectively a test for any model of selection, while the latter is a test for a specific model, e.g., the completely recessive model. The high uncertainty of the joint estimates of linkage and viability fitness may be related to their high statistical correlation (Eq. 3) and the oblique space of allowed segregation (Fig. 1). Also, the ability to detect viability loci is further reduced by the low allele frequency of their occurrence in populations. To increase the probability of detection, we have presented two experimental strategies: the use of several arrays of selfed progeny and the use of flanking markers. Several selfed progeny arrays, selected randomly from a population, increase the probability of detecting an allele with low frequency. Flanking markers increase the degrees of freedom for estimation and greatly reduce the sample size.

We have demonstrated that in our proposed design of multiple selfed progeny arrays, there exists an optimal number of selfed progeny arrays (assuming the total number of progeny assayed is fixed). Although this optimum specifically depends upon unknown parameter values, a general rule of thumb for an experiment of total size 2,000 is that 32 selfed progeny arrays are required if the selected allele frequency is as low as 0.03 and that 16 arrays are required if the allele frequency is higher than 0.15. These numbers roughly agree with the relationship $Pr(q)=1-[1-2q(1-q)]^N$, where $Pr(q)$ is the probability of detecting an allele in at least one of N arrays, which can be used as a simple rule to determine this experimental allocation.

We have also demonstrated that with two flanking markers, the required sample sizes are reduced by approximately 95% relative to the case of a single marker. In other words, the assay of a second, flanking marker is much more efficient than the assay of another progeny (for a single marker). With two markers, the degrees of freedom in the data are much greater, and the statistical correlation between *c* and *s* reduced, resulting in the great increase of power. Although it may be difficult to locate a second, linked marker, the increased efficiency of flanking markers places the highest priority upon an attempt to find linked markers for the study of loci controlling inbreeding depression.

Studies of inbreeding depression in even small portions of a genome, via the routine assay of a few isozyme markers, would provide valuable information about viability loci (Fu and Ritland 1994). However, a more "saturated" marker map (e.g., Tanksley et al. 1992) is desirable not only to increase the span of the genome surveyed, but also to increase the probability that a locus causing inbreeding depression is flanked by two informative markers. If experiments involving saturated marker maps are beyond the resources of an investigator, the best route is to survey small portions of the genome with pairs of linked markers. However, linked isozyme loci may not be easily found, given the relatively small number (10–20) of isozyme loci that can be routinely assayed in organisms. Although RAPD (randomly amplified polymorphic DNA, Williams et al. 1990) markers are much more numerous and linkage easier to find, their dominance makes them inappropriate for the methods considered here. A new class of markers, microsatellites (Herne et al. 1992; Morgante and Olivieri 1993), seem to show greater promise for use as flanking, co-dominant markers for characterizing loci causing inbreeding depression.

Acknowledgements We thank S. C. H. Barrett, J. Dole, S. Graham, P.W. Hedrick, R. Latta, and J. Z. Lin for their comments on the earlier version of the manuscript. This work was supported by a University of Toronto Open Doctoral Fellowship and a Mary H. Beatly Fellowship to YBF, and an operating grant from NSERC to KR.

References

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Carbonell EA, Asins MJ, Baselga M, Balansard E, Gerig TM (1993) Power studies in the estimation of genetic parameters and the localization of quantitative trait loci for backcross and doubled haploid populations. Theor Appl Genet 86:411–416

Charlesworth B (1992) Evolutionary rates in partially self-fertilization species. Am Nat 140:26–148

Charlesworth D, Charlesworth B (1987) Inbreeding depression and its evolutionary consequences. Annu Rev Ecol Syst 18:237–268

Darvasi A, Weinreb A, Minke V, Weller JI, Soller M (1993) Detecting marker-QTL-linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134:943–951

Fu YB, Ritland K (1994) Evidence for the partial dominance of viability genes contributing to inbreeding depression in *Mimulus guttatus*. Genetics 136:323–331

Grant V (1975) Genetics of flowering plants. Columbia University Press, New York

Hedrick PW, Muona O (1990) Linkage of viability genes to marker loci in selfing organisms. Heredity 64:67–72

Herne CM, Gosh S, Todd JA (1992) Microsatellites for linkage analysis of genetic traits. Trends Genet 8:288–294

Knapp SJ, Bridges WC, Birkes D (1990) Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79:583–392

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lyttle TW (1991) Segregation distorters. Annu Rev Genet 25:511–557

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J 3:175–182

Savolainen O, Kärkkäinen K, Kuittinen H (1992) Estimating number of embryonic lethals in conifers. Heredity 69:308–314

Simmons MJ, Crow JF (1977) Mutations affecting fitness in *Drosophila* populations. Annu Rev Genet 11:49–78

Soller M, Brody T, Genizi A (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47:35–39

Sorensen FC (1967) Linkage between marker genes and embryonic lethal factors may cause disturbed segregation ratios. Silvae Genet 16:132–134

Stuart A, Ord JK (1991) Kendall's advanced theory of statistics, vol. 2: classical inference and relationship. Edward Arnold, London

Tanksley SD, Ganal MW, Prince JP, De Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB, Messeguer R, Miller JC, Miller L, Paterson AH, Pineda O, Roder MS, Wing RA, Wu W, Young ND (1992) High density molecular linkage maps of the tomato and potato genomes. Genetics 132:1141–1160

Van der Beek S, Van Arendonk JAM (1993) Criteria to optimize designs for detection and estimation of linkage between marker loci from segregating populations containing several families. Theor Appl Genet 86:269–280

Wagner H, Weber WE, Wricke G (1992) Estimating linkage relationship of isozyme markers and morphological markers in sugar beet (*Beta vulgaris* L.) including families with distorted segregations. Plant Breed 108:89–96

Weeden NF, Wendel JF (1990) Genetics of plant isozymes. In: Soltis DE, Soltis PS (eds) Isozymes in plant biology. Dioscorides Press, Ore., USA, pp 46–72

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Zamir D, Tadmor Y (1986) Unequal segregation of nuclear genes in plants. Bot Gaz 147:355–358