Heredity **64** (1990) 67–72

Received 21 June 1989

Linkage of viability genes to marker loci in selfing organisms

Philip W. Hedrick* and
Outi Muona†

* Department of Biology, Pennsylvania State University, University Park, PA 16802.
† Department of Genetics, University of Oulu, SF-90570 Oulu, Finland.

Methodology for determining the linkage and effect of viability genes in selfing organisms (both intragametophytic and regular selfing) is developed. The maximum likelihood estimate of the recombination fraction and the selective effect is determined using a progeny array from a heterozygous parent. The method of lod scores, commonly used in human genetics, is applied to this situation. An example from Scots pine is given and the effect of polyembryony and segregation distortion are discussed.

INTRODUCTION

Determining the linkage of loci affecting fitness to other loci is generally difficult. In some outbreeding organisms, such as in *Drosophila melanogaster*, the large number of known marker loci along with special breeding schemes can be used to localize fitness variants. In other organisms, linkage maps have been recently developed using a combination of RFLP and allozyme markers (*e.g.*, Paterson *et al.*, 1988), making it feasible in the future to identity fitness variants.

The techniques developed to identify quantitative trait loci by Lander and Botstein (1989) utilizes backcrosses, *i.e.*, two generations of breeding, and does not estimate the effect of the loci. In plants with either a long generation length, *e.g.*, pines, or which are relatively difficult to cross, *e.g.*, homosporous ferns, the backcross approach may not generally be feasible. In addition, it would be useful to determine the extent of the effect of the variant on viability as well as its location. Therefore, here we present techniques that can be used in selfing organisms to determine the effect of viability variants and their linkage to codominant marker loci, *e.g.*, allozyme variants or RFLPs. First, we will discuss estimation in organisms that have intragametophytic selfing, *e.g.*, homosporous ferns, the most extreme form of selfing (*e.g.*, Hedrick, 1987). Second, we will discuss estimation for organisms with regular (intergametophytic) selfing.

INTRAGAMETOPHYTIC SELFING

One marker locus

Let us assume that a heterozygote, say A_1A_2 for codominant locus A produces an array of progeny by intragametophytic selfing, *i.e.*, all progeny are either homozygous A_1A_1 or homozygous A_2A_2. Assume that linked to locus A is a locus with an allele influencing viability and that the amount of recombination between the loci is c. When both loci are heterozygous, then the genotype will be either A_1l/A_2+ or A_1+/A_2l where l indicates the deleterious allele and + indicates the wild type allele. In the following, we will assume that the viability variant is on the same chromosome as A_1. If it is linked to A_2, then similar predictions ensue. Since it is not known to which A allele the viability variant is linked, strictly one should examine both possibilities. However, as in the Scots pine example below, one alternative is generally orders of magnitude more likely than others.

Given intragametophytic selfing of genotype A_1l/A_2+, there are four possible progeny types (table 1). If we assume that l is a recessive viability variant, then the frequencies of A_1A_1 and A_2A_2 progeny are

$$P_{11}=\frac{1-s+sc}{2-s}$$
$$P_{22}=\frac{1-sc}{2-s}. \qquad (1)$$

Table 1 The expected progeny genotypes and their frequencies from an A_1l_1/A_2+ individual when there is a linked viability allele and intragametophytic selfing

Progeny genotypes	General	Frequency		Fitness
		$c=\frac{1}{2}$	$c=0$	
A_1l/A_1l	$\frac{1-c}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$1-s$
A_2+/A_2+	$\frac{1-c}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	1
A_1+/A_1+	$\frac{c}{2}$	$\frac{1}{4}$	0	1
A_2l/A_2l	$\frac{c}{2}$	$\frac{1}{4}$	0	$1-s$

Note that only $(2-s)/2$ of the progeny survive because of this variant. If the variant is a recessive lethal ($s=1$), then $p_{11}=c$, $p_{22}=1-c$, and only $\frac{1}{2}$ the progeny survive. Because in this case there is only one independent equation with two unknowns, s and c, the same P_{11} and P_{22} values can result from different combinations of recombination and selection. For example $P_{11}=0{\cdot}33$ when $c=0{\cdot}0$ and $s=0{\cdot}5$, $c=0{\cdot}22$ and $s=0{\cdot}75$, and $c=0{\cdot}33$ and $s=1{\cdot}0$ as shown in fig. 1 by closed circles.

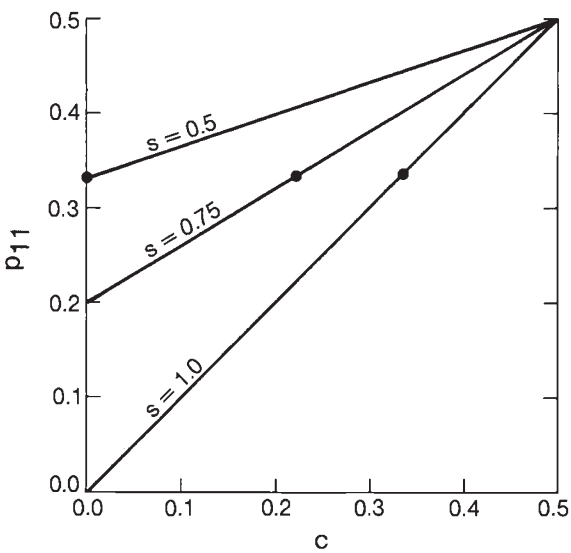

Figure 1 The expected frequency of genotype A_1A_1 (P_{11}) in progeny produced by intragametophytic selfing of a A_1A_2 individual when there is a lethal ($s=1$) or a deleterious ($s=0{\cdot}75$, $0{\cdot}5$) allele linked to A_1 (c is the rate of recombination between the loci).

The probability of a progeny or family (F) array assuming c recombination between the loci and with N total progeny, N_{11} individuals of genotype A_1A_1 and N_{22} of genotype A_2A_2, is

$$\Pr(F/c)=\frac{N!}{N_{11}!N_{22}!}P_{11}^{N_{11}}P_{22}^{N_{22}}. \qquad (2)$$

The ratio of the probability of a given family array with c recombination to that with no linkage, $c=\frac{1}{2}$, is known as relative probability and the $\log_{10}$ of the relative probability is known as the lod score (*e.g.*, Haldane and Smith, 1947; Morton, 1955; Ott, 1985). A general form for the lod score is

$$z=\sum_i N_i \log_{10}\frac{P_i}{P_i^*} \qquad (3)$$

where P_i and P_i^* are the proportions expected for the ith genotypic progeny class with recombination c and no linkage, respectively, and N_i is the number observed in the ith class. The maximum likelihood estimates for the various combinations of selfing for viability variants are the values of recombination and selection for which the lod score is a maximum. In general, a lod score of three or more is the minimum acceptable value to demonstrate significant linkage between two loci (Conneally *et al.*, 1985). A simple approach to determine the 95 per cent confidence interval for an estimate is to determine the parameter values one lod score value below the maximum lod score given that the maximum is three or greater.

The observed N_i values can be used in expression (3) to give a joint estimate of the recombination and selection levels. In general it is not possible in this case to determine what combination of s and c would best explain given N_i values. However, if a low proportion of progeny survive, then this may indicate higher s and higher c values rather than lower s and lower c values.

Two marker loci

When two linked marker loci are available, it is possible to distinguish between a loosely linked lethal (or near lethal) and a tightly linked detrimental allele. Let us assume that there are two marker loci linked to a viability locus so that the parental plant is the multiple heterozygote, A_1lB_1/A_2+B_2, where the viability locus is between the two markers and c_1 and c_2 are the probabilities of recombination between locus A or locus B, respectively, and the viability variant.

Assuming, intragametophytic selfing of this genotype and no recombinational interference, the

Table 2 The expected progeny genotypes from an $A_1A_2B_1B_2l+$ individual with intragametophytic selfing for the three possible gene orders given A_1, B_1, and l are on the same chromosome

	Frequency			
Progeny genotypes	A − *l* − B	*l* − A − B	A − B − *l*	Fitness
$A_1A_1B_1B_1ll$	$\frac{(1-c_1)(1-c_2)}{2}$	$\frac{(1-c_1)(1-c_2)}{2}$	$\frac{(1-c_1)(1-c_2)}{2}$	$1-s$
$A_2A_2B_2B_2{+}{+}$	$\frac{(1-c_1)(1-c_2)}{2}$	$\frac{(1-c_1)(1-c_2)}{2}$	$\frac{(1-c_1)(1-c_2)}{2}$	1
$A_1A_1B_2B_2{+}{+}$	$\frac{c_1(1-c_2)}{2}$	$\frac{c_1c_2}{2}$	$\frac{c_1(1-c_2)}{2}$	1
$A_2A_2B_1B_1ll$	$\frac{c_1(1-c_2)}{2}$	$\frac{c_1c_2}{2}$	$\frac{c_1(1-c_2)}{2}$	$1-s$
$A_1A_1B_2B_2ll$	$\frac{(1-c_1)c_2}{2}$	$\frac{(1-c_1)c_2}{2}$	$\frac{c_1c_2}{2}$	$1-s$
$A_2A_2B_1B_1{+}{+}$	$\frac{(1-c_1)c_2}{2}$	$\frac{(1-c_1)c_2}{2}$	$\frac{c_1c_2}{2}$	1
$A_2A_2B_2B_2{+}{+}$	$\frac{c_1c_2}{2}$	$\frac{c_1(1-c_2)}{2}$	$\frac{(1-c_1)c_2}{2}$	1
$A_2A_2B_2B_2ll$	$\frac{c_1c_2}{2}$	$\frac{c_1(1-c_2)}{2}$	$\frac{(1-c_1)c_2}{2}$	$1-s$

progeny genotypes are given in the second column of table 2. The expected frequencies of the four homozygotes after weighting by their viabilities are

$$\begin{aligned}
P(A_1A_1B_1B_1) &= \frac{c_1c_2+(1-c_1)(1-c_2)(1-s)}{2-s}\\
P(A_1A_1B_2B_2) &= \frac{c_1(1-c_2)+(1-c_1)c_2(1-s)}{2-s}\\
P(A_2A_2B_1B_1) &= \frac{(1-c_1)c_2+c_1(1-c_2)(1-s)}{2-s}\\
P(A_2A_2B_2B_2) &= \frac{(1-c_1)(1-c_2)+c_1c_2(1-s)}{2-s}
\end{aligned} \qquad (4)$$

Note that we now have three independent equations and three parameters to be estimated. The P_i values of expression (4) are then calculated for various c_1, c_2, and s values and P_i^* values calculated for $c_1 = c_2 = 0{\cdot}5$. Using the observed numbers of the four homozygotes, a lod score can then be calculated. The same procedure should also be carried out for the other gene orders whose progeny arrays are given in columns three and four of table 2.

For example, let us assume that the numbers observed of genotypes $A_1A_1B_1B_1$, $A_1A_1B_2B_2$, $A_2A_2B_1B_1$, and $A_2A_2B_2B_2$ in a progeny array are 5, 13, 4, and 26, respectively. For these numbers and gene order A − *l* − B, the maximum lod score is 5·373 when $c_1 = 0{\cdot}33$, $c_2 = 0{\cdot}07$, and $s = 0{\cdot}84$ (the other gene orders give much lower numbers). In other words, there is a nearly lethal allele seven map units from marker locus B and between the two markers.

REGULAR SELFING

Let us assume that a heterozygote for a codominant marker locus reproduces by selfing and that linked to it is a locus affecting viability. If the heterozygote is A_1l/A_2+, then the progeny genotypes and their fitnesses are given in table 3. Assuming that l is deleterious recessive, then the frequencies of the genotypes A_1A_1, A_1A_2, and AA_2 are

$$\begin{aligned}
P_{11} &= \frac{1-s+2sc-sc^2}{4-s}\\
P_{12} &= \frac{2(1-sc+sc^2)}{4-s}\\
P_{22} &= \frac{1-sc^2}{4-s}
\end{aligned} \qquad (5)$$

and $(4-s)/4$ of the progeny survive. If $s = 1$, then $\frac{3}{4}$ of the progeny survive. When a lethal is

Table 3 The expected progeny types and their frequencies from an A_1l/A_2+ individual when there is self-fertilization

Progeny genotypes	General	Frequency		Fitness
		$c=\frac{1}{2}$	$c=0$	
A_1l/A_1l	$(1-c)^2/4$	$\frac{1}{16}$	$\frac{1}{4}$	$1-s$
A_1l/A_2+	$(1-c)^2/2$	$\frac{1}{8}$	$\frac{1}{2}$	1
A_2+/A_2+	$(1-c)^2/4$	$\frac{1}{16}$	$\frac{1}{4}$	1
A_1l/A_1+	$c(1-c)/2$	$\frac{1}{8}$	0	1
A_1l/A_2l	$c(1-c)/2$	$\frac{1}{8}$	0	$1-s$
A_2+/A_1+	$c(1-c)/2$	$\frac{1}{8}$	0	1
A_2+/A_2l	$c(1-c)/2$	$\frac{1}{8}$	0	1
A_1+/A_1+	$c^2/4$	$\frac{1}{16}$	0	1
A_1+/A_2l	$c^2/2$	$\frac{1}{8}$	0	1
A_2l/A_2l	$c^2/4$	$\frac{1}{16}$	0	$1-s$

completely linked to the marker ($c=0$), then the frequencies of the three progeny genotypes A_1A_1, A_1A_2, and A_2A_2 after selection are 0, $\frac{2}{3}$, and $\frac{1}{3}$, respectively. Sorensen (1967) in fact examined a linked recessive lethal model as an explanation for segregation ratios differing from 3:1 expectations for recessive traits. The case he considered corresponds to the case here where there are only two types of progeny having frequencies P_{11} and $P_{12}+P_{22}$.

With regular selfing, there are two independent equations with two unknowns so that we can obtain a single estimate of s and c that maximizes the lod score. In other words, because there are three genotypic classes, we can estimate both parameters with only one locus.

As an example, self-fertilization of a Scots pine heterozygous for fluorescent esterase gave a progeny array of $N_{11}=6$, $N_{12}=45$ and $N_{22}=24$. Examining the combination of possible values of s and c, the maximum lod score is 3·167 when $c=0{\cdot}084$ and $s=0{\cdot}896$. We should note that if we assume that the allele is a recessive lethal $s=1$, that the maximum lod score of 3·162 occurs when $c=0{\cdot}127$.This viability variant is the third locus on linkage group B in Scots pine (Szmidt *et al.*, 1984). Fig. 2 gives the 95 per cent confidence interval for this estimate, *i.e.*, the combinations of c and s that give a lod score of 2·167.

The megagametophytic tissue has the same haploid genotype as the female gamete. Examining this tissue for the 75 progeny showed that 28 and 47 gametes were A_1 and A_2, respectively. The inferred male gametic contribution is then 29 A_1 and 46 A_2 gametes, very similar numbers to that

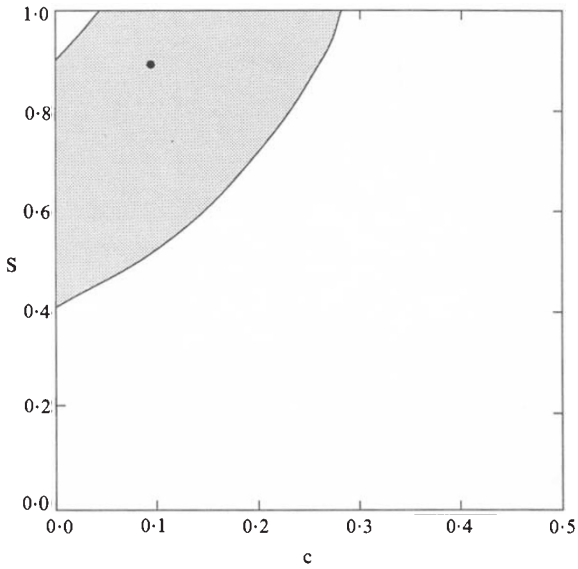

Figure 2 The 95 per cent confidence interval (between the two lines) of c and s for an allele linked to fluorescent esterase in Scots pine.

of the female gamete. Because of this similarity, the observations are consistent with an embryonic recessive detrimental lethal linked to A_1. Note that such a lethal would reduce the number of A_1 female gametes as observed (*e.g.*, Cheliak *et al.*, 1984; Strauss and Conkle, 1986).

DISCUSSION

The use of biochemical markers to map loci influencing quantitative traits or fitness characteristics in a variety of organisms promises to bring a new day to studies of evolutionary genetics. Here we have shown how progeny arrays from selfing species can be used to determine the effect of viability variants and their linkage to marker loci. Utilizing these techniques, we have identified a near lethal allele at a locus closely linked to an esterase marker in Scots pine. As the genomes of selfing species are saturated with markers and progeny arrays are generated, many other viability variants should similarly be documented.

Segregation distortion variants

There are several reports of what appear to be non-Mendelian segregation ratios for allozyme loci in pines. When there is intragametophytic selfing, two loci are again necessary to estimate the segregation ratio m (the proportion of m alleles from

$a+m$ heterozygote) and c. For regular selfing, let us assume that the parental genotype is A_1+/A_2m and that the segregation distortion occurs only in one sex (*e.g.*, Hartl, 1977). Using the same approach as earlier, then the frequencies of genotypes A_1A_1, A_1A_2, and A_2A_2 are

$$P_{11}=\tfrac{1}{2}(1-c-m+2cm)$$
$$P_{12}=\tfrac{1}{2} \qquad (6)$$
$$P_{22}=\tfrac{1}{2}(c+m-2cm).$$

Notice that these equations give proportions that are somewhat different than those given in equation (5) for a linked viability allele. For example, if $c=0{\cdot}1$ and $s=1$ then $P_{11}=0{\cdot}063$, $P_{12}=0{\cdot}067$, and $P_{22}=0{\cdot}33$ while if $m=0{\cdot}874$, and $c=0$, then $P_{11}=0{\cdot}063$, $P_{12}=0{\cdot}5$, and $P_{22}=0{\cdot}437$. In other words for a given P_{11} value, there is a higher proportion of heterozygotes and lower proportion of A_2A_2 for the linked lethal than for the segregation distortion allele. However, the maximum likelihood values of c and m cannot be estimated using lod scores because (6) has only one independent equation with two unknowns. In other words, one would need to examine two linked marker loci in order to estimate c and m. Furthermore, if megagametophytic tissue is analyzed as was in the Scots pine example above, segregation distortion may be eliminated as a factor because it occurs in only one sex.

Polyembryony

Many plants have polyembryony so that, for example in gymnosperms each ovule contains several archaegonia (*e.g.*, Sorensen, 1982), and in some ferns each gametophyte contains several archaegonia (*e.g.*, Klekowski, 1982). Usually only one embryo develops to form the mature embryo and when embryos are homozygous for recessive lethals, they do not develop. To illustrate how this may affect the expected proportions of the different genotypes, let us examine the simplest case, with two independent embryos and intragametophytic selfing (table 4). Again, the expected proportions of A_1A_1 and A_1A_2 genotypes when there is a linked lethal are c and $1-c$, as when there is only one embryo. However, instead of $\frac{1}{2}$ the ovules producing embryos, $\frac{3}{4}$ of the ovules produce embryos. The same result is true for regular selfing, *i.e.*, the proportion of the genotyupes at the marker locus remains unchanged with polyembryony, but the proportion of ovules producing embryos is higher than when there is only one embryo.

Table 4 The embryonic genotypes when there are two independent embryos and intragametophytic selfing

Embryos	Frequency	Genotype at locus A
$A_1l/A_1l, A_1l/A_1l$	$\frac{(1-c)^2}{4}$	—
$A_1l/A_1l, A_2+/A_2+$	$\frac{(1-c)^2}{2}$	A_2A_2
$A_1l/A_1l, A_1+/A_1+$	$\frac{c(1-c)}{2}$	A_1A_1
$A_1l/A_1l, A_2l/A_2l$	$\frac{c(1-c)}{2}$	—
$A_2+/A_2+, A_2+/A_2+$	$\frac{(1-c)^2}{4}$	A_2A_2
$A_2+/A_2+, A_1+/A_1+$	$\frac{(1-c)^2}{2}$	$\frac{1}{2}A_1A_1, \frac{1}{2}A_2A_2$
$A_2+/A_2+, A_2l/A_2l$	$\frac{c(1-c)}{2}$	A_2A_2
$A_1+/A_1+, A_1+/A_1+$	$\frac{c^2}{4}$	A_1A_1
$A_1+/A_1+, A_2l/A_2l$	$\frac{c^2}{2}$	A_1A_1
$A_2l/A_2l, A_2l/A_2l$	$\frac{c^2}{4}$	—

— No viable progeny

Acknowledgements We thank Dr Veikko Koski of the Finnish Forest Research Institute for providing the selfed seed used in the example. O.M. acknowledges financial support from the Academy of Finland. We thank an anonymous reviewer for comments that added greatly to our final version and to comments by J. Ott and F. Sorensen on an earlier version.

REFERENCES

CHELIAK, W. M., MORGAN, K., DANCIK, B. P., STROBECK, C. AND YEH, F. C. H. 1984. Segregation of allozymes in megagametophytes viable seeds from a natural population of jack pine, *Pinus banksiana* Lamb. *Theoret. Appl. Genet.*, *40*, 356-359.

CONNEALLY, J. H. EDWARDS, J. H., KIDD, K. K., LALOUEL, J.-M., MORTON, N. E., OTT, J. AND WHITE, R. 1985. Report of the committee on methods of linkage analysis and reporting. *Cytogenet. Cell Biol.*, *40*, 356-359.

HALDANE, J. B. S. AND SMITH, C. A. B. 1947. A new estimate of the linkage between the genes for color blindness and hemophilia in man. *Ann. Eugen.*, *14*, 10-31.

HARTL, D. L. 1977. Applications of meiotic drive in animal breeding and population control, in Kempthorne, E. *et al.* (eds) *Proc. Intern. Conf. Quant. Genet.*, Iowa State University Press, Ames, IA, pp. 63-88.

HEDRICK, P. W. 1987. Population genetics of intragametophytic selfing. *Evolution*, *41*, 137-144.

KLEKOWSKI, E. J. 1982. Genetic load and soft selection in ferns. *Heredity, 49*, 191-197.

LANDER, E. S., AND BOTSTEIN, A. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics, 121*, 185-189.

MORTON, N. E. 1955. Sequential tests for the detection of linkage. *Amer. J. Hum. Genet., 7*, 277-318.

OTT, J. 1985. *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore, MD.

PATERSON, A. H., LANDER, E. S., HEWITT, J. D., PETERSON, S., LINCOLN, S. E. AND TANKSLEY, S. D. 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphism. *Nature 335*, 721-726.

SORENSEN, F. C. 1967. Linkage between marker genes and embryonic lethal factors may cause disturbed segregation ratios. *Silvae Genet., 16*, 132-134.

SORENSEN, F. C. 1982. The role of polyembryony and embryo viability in the genetic system of confers. *Evolution, 36*, 725-733.

STRAUSS, S. H. AND CONKLE, M. T. 1986. Segregation, linkage, and diversity of allozymes in knobcone pine. *Theoret. Appl. Genet., 72*, 483-493.

SZMIDT, A. E., MUONA, O. AND YAZDANI, R. 1984. Linkage relationships in Scots pine (*Pinus Sylvestris* L.). In *Genetics Studies of Scots Pine* (Pinus sylvestris L.) *Domestication by Means of Isozyme Analysis.* Univ. Agr. Sci., Umea, Sweden.