Theor Appl Genet (1988) 76:125–128

The detection and estimation of linkage using doubled haploid or single seed descent populations

J. W. Snape
AFRC Institute of Plant Science Research, Cambridge Laboratory, Trumpington, Cambridge, UK

Received October 19, 1987; Accepted February 12, 1988
Communicated by G. Wenzel

Summary. In many plant species, particularly those of agricultural importance, there is now much effort being devoted to developing comprehensive genetic maps using biochemical and molecular markers. Because these techniques often involve destructive sampling of individual plants the use is increasingly made of homozygous or near-homozygous recombinant lines for linkage studies in preference to F_2 or backcross generations. The present paper describes methods for the detection and estimation of linkage using such generations for commonly encountered genetic situations.

Key words: Linkage – Haploids – Single seed descent

Introduction

Linkage mapping in plants has conventionally used generations derived from the F_1 between homozygous parents differing at appropriate loci – either the first backcrosses or the F_2. Not infrequently, selfed seed of these generations is also assessed to progeny test the preceding individual plants to obtain a fuller classification of genotypes. Most genetic maps have been built up using such procedures through assessing variation for morphological characters, either natural or induced (including those for disease resistance), and, more recently, electrophoretic markers. Indeed, biochemical and molecular markers are now greatly expanding the genetic maps of many species, for example in bread wheat, *Triticum aestivum*, about 40% of the loci presently mapped are those for isozyme and protein variation (Chenicek and Hart 1987). This proportion is likely to increase particularly with the advent of data on restriction fragment length polymorphisms (RFLP's).

In conventional mapping studies the individual plant is evaluated as the unit of segregation. However, as the number of loci being concurrently assessed increases and as destructive sampling of plant tissue is often necessary this can create practical difficulties. Although selfed seed from the backcross, or the F_3, can compensate for this, progeny testing can also introduce sampling problems which can complicate genotypic classification. An alternative strategy which can resolve these difficulties is to use random near-homozygous or completely homozygous lines derived by single seed descent (Brim 1966) or doubled haploid methods. Within individual lines, plants are genetically uniform and therefore different tissues can be used for destructive sampling – embryos, endosperm, leaves, spikes, whilst the allelic classifications all relate to the same genotype. A further advantage is that the number of alternative genotypes in these generations per locus is less since few or no heterozygotes are present. For example if n loci are segregating then the F_∞ population will contain 2^n different genotypes compared to 3^n in a F_2.

Because of the increasing use of this type of material, the present paper presents methods for the detection and estimation of linkage in populations of random homozygous lines derived by doubled haploid methods or single seed descent for commonly encountered genetical situations.

Detection and estimation of linkage using doubled haploid (DH) lines

Using appropriate techniques DH's can be produced from any filial generation (Snape and Simpson 1981).

However, the most common population produced is that from the F_1 between two inbred lines and it is this generation that is most suitable for mapping studies. The great advantage of DH genotypes is, of course, that the dominance relationships at any locus are irrelevant since only homozygous progenies are present and thus, a complete classification of genotypes is always possible. In effect it is the population of gametes from the parent F_1 which is being assessed. Consequently for a cross segregating for n loci then 2^n genotype classes should be present in the recombinant population. In the absence of linkage then each class is expected at a frequency of $(\frac{1}{2})^n$.

If segregation for two loci, A/a, B/b, linked with a recombination frequency, p, is considered, then the frequency of the four possible genotypes in a population of F_1 (AaBb) derived DH lines are:

DH Genotype	Alleles in parents in:		Number observed
	Coupling	Repulsion	
AABB	$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	a
AAbb	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$	b
aaBB	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$	c
aabb	$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	d

These expectations are, of course, analogous to those of a backcross to the double recessive parent.

In the absence of linkage, with complete genotypic manifestation, and no epistasis, then the four phenotype classes are expected with equal frequency. Thus the presence of linkage can be detected using a χ^2 test as described by Mather (1938) where 3 degrees of freedom can be partitioned into three orthogonal comparisons for segregation of the two individual loci and for linkage:

Observed	a	b	c	d	total n
Expected	n/4	n/4	n/4	n/4	total n

If linkage is present then maximum likelihood estimate of p can be obtained from the equations:

alleles in coupling phase: $p=\frac{b+c}{n}$

alleles in repulsion phase: $p=\frac{a+d}{n}$

The variance of p, Vp, is calculated from $p(1-p)/n$ in the usual way (Mather 1938).

For the segregation of three loci, A/b, B/b, C/c eight genotypic classes are expected. Considering recombination frequencies p_1 and p_2 between A/a-B/b, and B/b-C/c respectively, with all alleles in coupling then expectations are:

Genotype	Frequency	Observed numbers
AABBCC	$\frac{1}{2}(1-p_1)\cdot(1-p_2)$	a
AABBcc	$\frac{1}{2}(1-p_1)\cdot p_2$	b
AAbbCC	$\frac{1}{2}p_1\cdot p_2$	c
AAbbcc	$\frac{1}{2}p_1\cdot(1-p_2)$	d
aaBBCC	$\frac{1}{2}p_1\cdot(1-p_2)$	e
aaBBcc	$\frac{1}{2}p_1\cdot p_2$	f
aabbCC	$\frac{1}{2}(1-p_1)p_2$	g
aabbcc	$\frac{1}{2}(1-p_1)(1-p_2)$	h
		n

Maximum likelihood equations for estimating the recombination frequencies are:

$$p_1=\frac{c+d+e+f}{n}$$

$$p_2=\frac{b+d+e+g}{n}$$

with variances of $\frac{p_1(1-p_1)}{n}$ and $\frac{p_2(1-p_2)}{n}$ respectively.

These are, again, analogous to the standard expectations for a backcross as given by Mather (1938). The equations can be adjusted appropriately for situations in which two alleles are in association and one in dispersion.

For situations where four or more alleles are linked on the same chromosome the same principles apply. However the computations become more complicated and computer programmes, such as that described by Gale et al. (1983), can be employed to produce maximum likelihood estimates of the appropriate recombination coefficients and also to evaluate the most likely gene order.

Detection and estimation of linkage using single seed descent (SSD) lines

The commonly practised form of SSD is to develop lines by inbreeding an F_2 population developed from two inbred parents. From a sample of F_2 individuals a single seed is taken from each to form the F_3 SSD population. This process is repeated in subsequent generations until a required level of homozygosity is achieved. Thus in each generation each individual plant grown is a representative of a different, ancestral F_2 plant. At the final generation of SSD all seed on each plant is harvested and these families form the nuclear seed of the SSD lines for assessment. There is no prescribed number of generations of SSD before lines are extracted although, commonly, the F_6 or F_7 are used (Snape et al. 1985), that is, 3 or 4

generations of SSD. It should be noted, however, that for linkage estimation the F_n genotype frequencies are usually established from assessment of their F_{n+1} or further bulked selfed generations.

If two loci A/a, B/b are again considered, then, in the absence of linkage, the frequencies of the nine possible genotypes in the F_n SSD population can be calculated from the expressions:

Genotype	Frequencies
AABB, aabb, aaBB, AAbb	$\frac{1}{4}[1-\frac{1}{2}^{(n-1)}]^2$
AABb, aaBb, AaBB, Aabb	$(\frac{1}{2})^n - \frac{1}{2}^{(2n-1)}$
AaBb	$(\frac{1}{2})^{(2n-2)}$

If the alleles are co-dominant, as is the case with most isozyme or RFLP markers, then a complete classification of genotypes is possible and the eight degrees of freedom available can be partitioned into orthogonal comparisons which detect deviations from expected segregation frequencies of alleles at each of the loci separately and for linkage. As n becomes large the frequencies of the heterozygous classes becomes very small and the genotype frequencies for AABB, aabb, AAbb and aaBB approach $\frac{1}{4}$. For situations where one or both loci exhibit dominance then the expected phenotypic frequencies can be calculated and tested for different genetic models by using the above equations and summing over the appropriate genotype frequencies.

If the presence of linkage is detected then the expected frequencies of parental and recombinant genotypes in any generation for two loci linked by a recombination frequency p are given by the recurrence equations described by Haldane and Waddington (1931).

If the frequencies of genotypes in the F_n generation are:

Genotype	Expected frequency	Observed frequency
AABB, aabb	$:C_n$	d, e
AAbb, aaBB	$:D_n$	j, s
AABb, AaBB, Aabb, aaBb	$:E_n$	g, f, k, m
AB · ab	$:H_n$	h
Ab · aB	$:G_n$	i

then the frequencies in the F_{n+1} generation are

$$C_{n+1} = C_n + \tfrac{1}{2}E_n + \tfrac{1}{4}(1-p)^2 H_n + p^2 G_n$$

$$D_{n+1} = D_n + \tfrac{1}{2}E_n + \tfrac{1}{4}p^2 H_n + \tfrac{1}{4}(1-p)^2 G_n$$

$$E_{n+1} = \tfrac{1}{2}E_n + \tfrac{1}{4}(2p-2p^2)(H_n + G_n)$$

$$H_{n+1} = \tfrac{1}{2}(1-p)^2 H_n + \tfrac{1}{2}p^2 G_n$$

$$G_{n+1} = \tfrac{1}{2}H_n + \tfrac{1}{2}(1-p)^2 G_n$$

In most circumstances, with either dominant or co-dominant genetic markers, the coupling and repulsion double heterozygotes will not be distinguishable from one another and, therefore, nine rather than ten genotypic classes will be scored.

The calculation of the recombination frequency in a F_2 generation for the different genetical situations that are likely to be encountered are given in the comprehensive paper by Allard (1956). For example with two co-dominant genetic markers giving a complete classification of genotypes, apart from coupling and repulsion heterozygotes, the estimate of p is obtained by solving the equation:

$$(d+e)\cdot\left[\frac{2}{p}\right] + (f+g+k+m)\cdot\left[\frac{(1-2p)}{p(1-p)}\right] + (j+s)\cdot\left[\frac{2}{p-1}\right] + (h+i)\cdot\left[\frac{2(2p-1)}{1-2p+2p^2}\right] = 0$$

Since estimates of p must take a value between 0 and 0.5 a simple method of obtaining an estimate is to use a computer programme to compute and minimise this equation by iteration.

The estimate obtained has a variance of:

$$\left[\frac{p(1-p)(1-2p+2p^2)}{2(1-3p+3p^2)}\right]\Big/ n \qquad \text{(Allard 1956)}$$

For later generations an algebraic solution for p although theoretically possible is extremely laborious. Thus the use of appropriate computer software is necessary. For example Snape et al. (1985) calculated exact maximum likelihood recombination frequencies for a F_5 SSD generation using a general optimising routine (Ross 1980).

However approximate estimates of the recombination frequency can be obtained from the observed proportions of homozygous recombinant and homozygous parental genotypes alone, since the frequencies of heterozygous genotypes in advanced generations is small. For example, Fig. 1 shows the expected frequencies of these genotypes in F_5, F_6, F_7 and F_∞ generations for different values of *p*. From the observed proportions of these homozygous classes in a population it is thus possible to obtain an approximate estimate of the recombination frequency directly from this graph.

For each population two estimates can be obtained, one from the frequency of homozygous parentals and the other from the frequency of homozygous recombinants. The mean value should approximate the maximum likelihood value. For example, the complete data of Snape et al. (1985) for linkage in wheat between the loci *Rf3* and *Gli-Bl* on chromosome 1B give a maximum likelihood value for the recombination frequency of 0.221. In the combined population for the two crosses the propor-

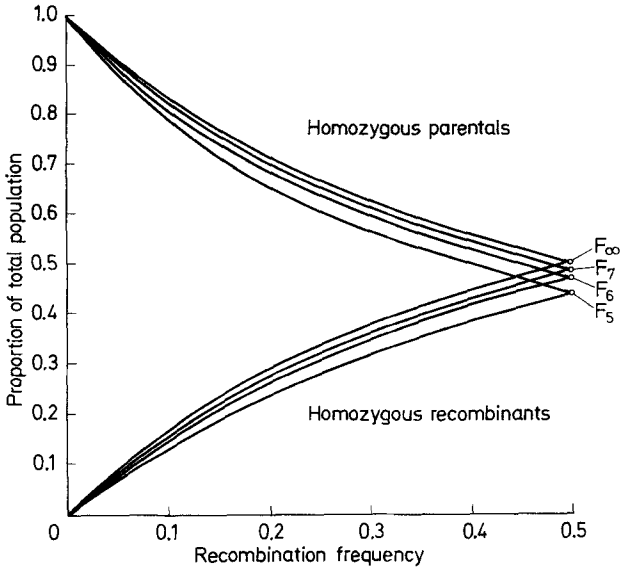

Fig. 1. The relationship between the proportions of homozygous parental and recombinant genotypes in F_5, F_6, F_7 and F_∞ single seed descent populations and the recombination frequency, p

tion of homozygous parental genotypes was 48/77 = 0.623 giving, from Fig. 1, an estimate of p of 0.23. The frequency of homozygous recombinant lines was 18/77 = 0.233 giving a value of 0.20. These values, with a mean of 0.215, approximate closely to the maximum likelihood value.

It is not possible to attach a standard error to these estimates using this method. However an approximate standard error can be obtained from assuming that the value of p is close to the F_∞ value and using the maximum likelihood variance of this.

In the F_∞ generation only the homozygous classes are present. These are expected in the frequencies:

$$C_\infty = \frac{1}{2(1+2p)} \quad \text{and} \quad D_\infty = \frac{p}{(1+2p)}$$

The maximum likelihood value of p can then be obtained from the observed frequencies as: $p = \frac{(j+s)}{(d+e)}$
This estimate has a variance of

$$\frac{p^2(1+2p)^2}{(j+s)(1+4p) - 4(d+e)p^2}$$

Thus an approximate error for values obtained from Fig. 1 can be obtained using this equation.

Discussion

Doubled haploid populations have both practical and computational advantages over SSD populations in calculating recombination frequencies when many loci are segregating. Unfortunately, however, they are technically difficult to produce and population sizes for most crosses are likely to be small. Because of these problems their use in mapping will probably be restricted to a few species and to particular genotypes. On the other hand, large populations of SSD lines are relatively easy and cheap to produce with most species and genotypes, and their use for mapping is likely to increase alongside the development of biochemical and molecular marker systems. It is not inconceivable that it will be possible to saturate genomes with markers separated by about 10 Cm (Gale, personal communication) in particular "model" crosses.

References

Allard RW (1956) Formulas and tables to facilitate the calculation of recombination values in Heredity. Hilgardia 24: 235–278

Brim CA (1966) A modified pedigree method of selection in soyabean. Crop Sci 6:220

Chenicek KJ, Hart GE (1987) Identification and chromosomal locations of aconitase gene loci in Triticeae. Theor Appl Genet 74:261–268

Gale MD, Law CN, Chojecki AJS, Kempton RA (1983) Genetic control of alpha-amylase production in wheat. Theor Appl Genet 64:309–316

Haldane JBS, Waddington CH (1931) Inbreeding and linkage. Genetics 16:357–374

Mather K (1938) The measurement of linkage in heredity. Methuen, London

Ross GJS (1980) Maximum likelyhood programme. Rothamstead Exp Statn. Harpenden UK

Snape JW, Simpson E (1981) The genetical expectations of doubled haploid lines derived from different filial generations. Theor Appl Genet 60:123–128

Snape JW, Flavell RB, O'Dell M, Hughes WG, Payne PI (1985) Intrachromosomal mapping of the nucleolar organiser region relative to three marker loci on chromosome 1B of wheat (*Triticum aestivum*). Theor Appl Genet 69:263–270