Theor Appl Genet (1992) 85:101–104

THEORETICAL AND APPLIED GENETICS

Theory for identification of marker locus-QTL associations in population by line crosses

J. W. Dudley

Department of Agronomy, University of Illinois, 1102 S. Goodwin Ave., Urbana, IL 61801, USA

Received December 2, 1991; Accepted March 24, 1992
Communicated by A. R. Hallauer

Summary. The objective of this paper is to present genetic theory demonstrating the conditions under which it should be possible to identify molecular marker-quantitative trait locus (QTL) associations in crosses of random-mating populations to inbreds. Using as an example the cross of a corn (*Zea mays* L.) population to an inbred, the expected disequilibrium for testcross and per se performance of F_2, F_3, BC_1 (to the inbred) and recombinant inbred generations was derived for cases where a marker allele is linked to an unfavorable QTL allele in the inbred and where the marker allele is linked to a favorable QTL allele in the inbred. Disequilibrium in segregating generations was shown to be a function of disequilibrium in the parent population, the frequency of marker and QTL alleles in the parent population, and the recombination distance between the marker and the QTL. To maximize the opportunity to identify a favorable QTL the following procedures are suggested:

(1) Select marker loci with alleles in the parent population which are not present in the inbred.
(2) Select populations known to have favorable QTL alleles not present in the inbred.
(3) Use as many marker loci as possible to enhance the probability of tight linkage between the marker and the QTL.

Key words: Marker assisted selection – RFLP – QTL – Quantitative genetics – Corn breeding

Introduction

The first use of marker-facilitated selection was described by Sax (1923). In recent years, the development of restriction fragment length polymorphism (RFLP) technology has provided a mechanism for the identification of loci (QTLs) affecting quantitative traits and the possibility of enhancing the effectiveness of marker-assisted selection (Stuber 1992). For corn breeding applications, either F_2, backcross, or random single-seed-descent inbred generations from crosses between homozygous lines (evaluated either per se or in testcrosses) have been used to identify marker locus-QTL associations (Stuber 1992). Use of these generations from bi-parental crosses assures linkage disequilibrium which is essential for identification of marker-QTL associations. However, in many corn breeding programs there are heterogenous populations (e.g., synthetics, open-pollinated varieties, exotic populations) from which it would be desirable to isolate genes for quantitative traits and incorporate them into elite germplasm. One method of utilizing such germplasm in corn breeding is to cross a population to an elite inbred and to select either in F_2 or backcross generations for individuals with desirable combinations of characteristics from the elite inbred and the donor population. As part of this process, the identification of molecular marker alleles associated with favorable QTL alleles in the population would be useful. The favorable QTL alleles could then be incorporated into elite germplasm by selecting for the appropriate marker alleles.

The objective of this paper is to evaluate, on a theoretical basis, the potential for identifying marker-QTL associations from population × inbred crosses and to discuss ways of utilizing these associations in breeding programs. The generations considered are the F_2, F_3, advanced random-mated generations from the F_2, backcross to the inbred, and recombinant inbreds derived from the F_2.

Gametic disequilibrium

To illustrate the problem, consider the following example. Dudley (1988) evaluated the potential of a number of corn populations for improving the hybrid Mo17 × B73. The population BS11(FR)C7 (subsequently BS11) was shown to have the greatest probability of improving this single cross. BS11 was also shown to be more closely related to Mo17 than B73 and thus should be a donor of useful alleles to Mo17. Consider the cross of BS11 × Mo17. BS11 is heterogenous and may not have marker loci in linkage equilibrium with a QTL. Mo17 can be considered homozygous. To identify a QTL from BS11 useful for improving Mo17, there must be linkage disequilibrium in the F_2 of the cross BS11 × Mo17. To evaluate the potential for linkage disequilibrium in a cross of this type, a model with two marker alleles per marker locus and two alleles per QTL was used.

Let the genotype of Mo17 be $M_1M_1t_1t_1$ where M_1 is a marker allele at marker locus 1 with an alternate allele m_1, t_1 is an unfavorable QTL allele at QTL1 with an alternate allele T_1. In BS11, let the frequency of M_1 be p and of m_1 be q and let the frequency of T_1 be w and of t_1 be x where $p+q=1$ and $w+x=1$. Assume BS11 is in random mating linkage equilibrium (this assumption is relaxed later). Then gamete frequencies in BS11 are:

$$M_1T_1 = p\,w;\quad M_1t_1 = p\,x;\quad m_1T_1 = q\,w;\quad \text{and}\quad m_1t_1 = q\,x.$$

The F_1 of the cross BS11 × Mo17 will have the following genotypic frequencies:

$$M_1T_1/M_1t_1 = p\,w;\qquad M_1t_1/M_1t_1 = p\,x;$$
$$m_1T_1/M_1t_1 = q\,w;\qquad m_1t_1/M_1t_1 = q\,x.$$

Gametes from the F_1 will be produced with the following frequencies:

$$M_1T_1 = 0.5\,p\,w + 0.5\,r\,q\,w;$$
$$M_1t_1 = 0.5\,p\,w + p\,x + 0.5\,(1-r)\,q\,w + 0.5\,q\,x;$$
$$m_1T_1 = 0.5\,(1-r)\,q\,w;$$
$$m_1t_1 = 0.5\,q\,x + 0.5\,r\,q\,w;$$

where r is the recombination value; $r=0$ for complete linkage between the marker locus and the QTL and $r=0.5$ for free recombination. If the F_1 were random-mated to linkage equilibrium, the gametic frequencies would be:

$$M_1T_1 = 0.25\,w\,(1+p);$$
$$M_1t_1 = 0.25\,(1+p)\,(2-w);$$
$$m_1T_1 = 0.25\,q\,w;$$
$$m_1t_1 = 0.25\,q\,(2-w).$$

Thus, the gametic disequilibrium contributed by crossing BS11 to Mo17 is $0.25\,q\,w\,(1-2\,r)$, which is 0 if there is no linkage ($r=0.5$) and maximum if $q=w=1$ and $r=0$, the case if the population is homozygous for m_1 and T_1 and there is no recombination. If $q=w=1$, the disequilibrium reduces to $0.25\,(1-2\,r)$, the value for the F_1 from a cross between two homozygous lines.

A second source of disequilibrium is that existing in BS11 prior to crossing to Mo17. If BS11 is not in equilibrium, then gametes from BS11 will have frequencies:

$$M_1T_1 = p\,w - D;\qquad M_1t_1 = p\,x + D;$$
$$m_1T_1 = q\,w + D;\qquad m_1t_1 = q\,x - D$$

where D is the gametic disequilibrium in BS11 and can be either positive or negative. A positive value of D means that the disequilibrium in BS11 changes gamete frequencies in the same direction as does disequilibrium resulting from linkage in the F_1. Assuming that BS11 is not in equilibrium, the total gametic disequilibrium is $0.25\,q\,w\,(1-2\,r)+0.5\,(1-r)\,D$. The first term is the disequilibrium contributed by the cross of BS11 to Mo17, and the second term is the additional disequilibrium contributed by the disequilibrium contained in BS11. Even if $r=0$, half the disequilibrium present in BS11 remains. If D is negative, then the total disequilibrium is reduced. Thus, maximum disequilibrium will be achieved when $r=0$; $q=w=1$; and D is positive.

In general terms, the gametic frequencies in the F_1 can be expressed as follows:

$$M_1T_1 = p'w' - D'$$
$$M_1t_1 = p'x' + D'$$
$$m_1T_1 = q'w' + D'$$
$$m_1t_1 = q'x' - D'$$

where p' = frequency of M_1, q' = frequency of m_1, w' = frequency of T_1, and x' = frequency of t_1 in the F_1 regardless of the frequency in the population. In the preceding example, $p'=0.5\,(1+p)$, $q'=0.5\,q$, and $D'=0.25\,q\,w\,(1-2\,r)+0.5\,(1-r)\,D$. Following t generations of random-mating the F_2, D_t (gametic disequilibrium in generation t) $=(1-r)^t\,D'$ (Li, 1955), and in the absence of selection or drift, $M_1T_1 = p'w' - (1-r)^t\,D'$; $M_1t_1 = p'x' + (1-r)^t\,D'$; $m_1T_1 = q'w' + (1-r)^t\,D'$; and $m_1t_1 = q'x' - (1-r)^t\,D'$.

The case just considered is one in which the favorable QTL allele is absent from Mo17. Disequilibrium can exist and marker-QTL associations can be identified if Mo17 contains the favorable QTL (i.e., has the genotype $M_1M_1T_1T_1$). In this case, general gametic frequencies are $M_1T_1 = p'w' + D'$, $M_1t_1 = p'x' - D'$, $m_1T_1 = q'w' - D'$, and $m_1t_1 = q'x' + D'$. Note that p' and q' are the same as before but $w'=0.5\,(1+w)$ and $x'=0.5\,x$. Disequilibrium created by crossing to the inbred will be $0.25\,q\,x\,(1-2\,r)$ and maximum disequilibrium will exist when $r=0$ and $q=x=1$.

Genotypic contrasts

Identification of marker allele-QTL associations requires differences in trait means between marker genotypes. The following common situations were considered: testcrosses of individual F_2 or BC_1 (to Mo17) plants to an inbred tester homozygous recessive for t_1t_1, per se performance of F_2 and backcross (to Mo17) plants, per se performance of F_3 lines produced by selfing individual F_2 plants, and both per se and testcross performance of recombinant inbred lines. Genotypic values for QTL genotypes are assigned as:

$$T_1T_1 = a;\qquad T_1t_1 = d;\qquad t_1t_1 = -a;$$

where a is half the difference between the homozygous genotypes and d is the coded value of the heterozygote ($d=a$ = complete dominance and $d=0$ = additivity). Frequencies of F_2 genotypes arising from random-mating F_1 plants, of backcross genotypes, and of recombinant inbred genotypes, were calculated using gametic frequencies which assume gametic disequilibrium in BS11. The frequency of recombinant inbred genotypes was calculated using the result given by Robbins (1918) for the frequency of genotypes resulting from selfing double heterozygotes when linkage is present.

For each type of progeny, genotypic values for individuals were then calculated. As an example, for the F_2 genotype M_1T_1/M_1t_1 the testcross genotype will be $0.5\,T_1t_1 + 0.5\,t_1t_1$ and the genotypic value is $0.5\,d - 0.5\,a$. For F_2 plants the genotype is T_1t_1 and the genotypic value is d. For F_3 progenies the genotype is $0.25\,T_1T_1 + 0.5\,T_1t_1 + 0.25\,t_1t_1$ and the genotypic value is 0.5 d. Based on frequencies of F_2 genotypes, marker genotype

means were calculated for all cases. As an example, F_2 testcross means (when $Mo17 = M_1 M_1 t_1 t_1$) are:

$$\overline{M_1 M_1} = 0.5\,[(w-2)\,a + w\,d] - [(a+d)/(1+p)]\,[0.5\,q\,w\,(1-2\,r) + (1-r)\,D]\,,$$

$$\overline{M_1 m_1} = 0.5\,[(w-2)\,a + w\,d] + [p\,(a+d)/q\,(1+p)]\,[0.5\,q\,w\,(1-2\,r) + (1-r)\,D]\,,$$

and

$$\overline{m_1 m_1} = 0.5\,[(w-2)\,a + w\,d] + [(a+d)/q]\,[0.5\,q\,w\,(1-2\,r) + (1-r)\,D]\,.$$

Genetic effects in the F_2 were evaluated by calculating α, the average effect of a marker allele substitution (Falconer, 1989) for F_2 testcrosses, F_2 plants and F_3 progenies. For backcross progenies, the marker genotypic contrast $M_1 M_1 - M_1 m_1$ was calculated for each case. The genotypic contrast $M_1 M_1 - m_1 m_1$ was calculated for recombinant inbreds. In the F_2, the frequencies of marker allele genotypes are:

$$M_1 M_1 = p'^2\,; \quad M_1 m_1 = 2\,p'q'\,; \quad \text{and} \quad m_1 m_1 = q'^2$$

for the general case.

In the general notation, $\alpha = p'\,\overline{M_1 M_1} - (p'-q')\,\overline{M_1 m_1} - q'\,\overline{m_1 m_1}$ where $\overline{M_1 M_1}$, $\overline{M_1 m_1}$ and $\overline{M_1 M_1}$ are genotypic means. In the specific cases of $Mo17 = M_1 M_1 t_1 t_1$ or $Mo17 = M_1 M_1 T_1 T_1$, the average effect of a marker allele substitution is:

$$\alpha = 0.5\,(1+p)\,\overline{M_1 M_1} - p\,\overline{M_1 m_1} - 0.5\,q\,\overline{m_1 m_1}\,.$$

Substituting expected values of marker genotypic means gives for F_2 testcross progenies:

$$\alpha\,(t_1) = -2\,[(a+d)/q\,(1+p)]\,[0.25\,q\,w\,(1-2\,r) + 0.5\,(1-r)\,D]$$

and

$$\alpha\,(T_1) = 2\,[(a+d)/q\,(1+p)]\,[0.25\,q\,x\,(1-2\,r) + 0.5\,(1-r)\,D]\,.$$

The designation (t_1) refers to the case were $Mo17 = M_1 M_1 t_1 t_1$. This convention will be used for all contrasts. For backcross testcross progenies, only one marker genotypic contrast is possible, $M_1 M_1 - M_1 m_1$. Under the assumptions and models used, the differences for the cases where $Mo17 = M_1 M_1 t_1 t_1$ and $Mo17 = M_1 M_1 T_1 T_1$ are:

$$\overline{M_1 M_1} - \overline{M_1 m_1}\,(t_1) = -2\,[(a+d)/q\,(1+p)]\,[0.25\,q\,w\,(1-2\,r) + 0.5\,(1-r)\,D]$$

and

$$\overline{M_1 M_1} - \overline{M_1 m_1}\,(T_1) = 2\,[(a+d)/q\,(1+p)]\,[0.25\,q\,x\,(1-2\,r) + 0.5\,(1-r)\,D].$$

which are the same as the average effects of marker allelic substitution in the F_2 [$\alpha\,(t_1)$ and $\alpha\,(T_1)$].

For F_2 per se data,

$$\alpha\,(t_1) = -4\,[(a+d\,x)/q\,(1+p)]\,[0.25\,q\,w\,(1-2\,r) + 0.5\,(1-r)\,D]$$

and

$$\alpha\,(T_1) = 4\,[(a+d\,x-d)/q\,(1+p)]\,[0.25\,q\,x\,(1-2\,r) + 0.5\,(1-r)\,D]\,.$$

For backcrosses per se data,

$$\overline{M_1 M_1} - \overline{M_1 m_1}\,(t_1) = -4\,[(a+d)/q\,(1+p)]\,[0.25\,q\,w\,(1-2\,r) + 0.5\,(1-r)\,D]$$

and

$$\overline{M_1 M_1} - \overline{M_1 m_1}\,(T_1) = 4\,[(a-d)/q\,(1+p)]\,[0.25\,q\,x\,(1-2\,r) + 0.5\,(1-r)\,D]\,.$$

For F_3 progenies,

$$\alpha\,(t_1) = -4\,[(a+0.5\,d\,x)/q\,(1+p)]\,[0.25\,q\,w\,(1-2\,r) + 0.5\,(1-r)\,D]$$

and

$$\alpha\,(T_1) = 4\,[(a+0.5\,d\,x-0.5\,d)/q\,(1+p)]\,[0.25\,q\,x\,(1-2\,r) + 0.5\,(1-r)\,D]\,.$$

The effect of using F_3 progenies is to reduce the contribution of the d effect by 0.5 compared to using F_2 plant data.

Using the result of Robbins (1918), the genotypic frequencies for recombinant inbreds were found to be:

$$M_1 M_1 T_1 T_1 = p'w' - D'\,[1 - r/(1+2\,r)]\,,$$

$$M_1 M_1 t_1 t_1 = p'x' + D'\,[1 - r/(1+2\,r)]\,,$$

$$m_1 m_1 T_1 T_1 = q'w' + D'\,[1 - r/(1+2\,r)]\,,$$

and

$$m_1 m_1 t_1 t_1 = q'x' - D'\,[1 - r/(1+2\,r)]\,.$$

For inbred per se data the contrasts are:

$$[\overline{M_1 M_1} - \overline{m_1 m_1}]\,(t_1) = -[8\,a/q\,(1+p)]\,[1 - r/(1+2\,r)] \cdot [0.25\,q\,w\,(1-2\,r) + 0.5\,(1-r)\,D]\,,$$

and

$$[\overline{M_1 M_1} - \overline{m_1 m_1}]\,(T_1) = [8\,a/q\,(1+p)]\,[1 - r/(1+2\,r)] \cdot [0.25\,q\,x\,(1-2\,r) + 0.5\,(1-r)\,D]\,.$$

For testcross data where the tester is $t_1 t_1$,

$$[\overline{M_1 M_1} - \overline{m_1 m_1}]\,(t_1) = -[(4\,(a+d)/q\,(1+p)]\,[1 - r/(1+2\,r)] \cdot [0.25\,q\,w\,(1-2\,r) + 0.5\,(1-r)\,D]\,,$$

$$[\overline{M_1 M_1} - \overline{m_1 m_1}]\,(T_1) = [4\,(a+d)/q\,(1+p)]\,[1 - r/(1+2\,r)] \cdot [0.25\,q\,x\,(1-2\,r) + 0.5\,(1-r)\,D]\,.$$

All contrasts, regardless of the type of data used, consist of two parts: one which is a function of the residual disequilibrium from the population, and one which is the linkage disequilibrium resulting from crossing the population to the inbred.

The effect of random-mating the F_2 prior to attempting to identify marker locus-QTL associations is to reduce α by a factor of $(1-r)^t$. Thus, if the F_2 were random-mated twice, $\alpha\,(t_1)$ for testcross progenies would be

$$-2\,(1-r)^2\,[(a+d)/q\,(1+p)]\,[0.25\,q\,w\,(1-2\,r) + 0.5\,(1+r)\,D]\,.$$

If $r = 0.1$, α after two generations of random-mating is 81% of α measured on F_2 plants. However, α is 98% of the F_2 value if $r = 0.01$. As r increases, the reduction in value of α increases. Thus, with tight linkage, advanced random-mated generations may be used to identify marker locus-QTL associations, but loosely linked associations will probably become undetectable.

Discussion

These results demonstrate that it should be possible to use population × inbred F_2, F_3, backcross, or recombinant inbred generations, to identify molecular marker-QTL associations. Individual plant data, testcross data, or line per se data can be used. Regardless of the type of data, two sources of gametic disequilibrium are present in such crosses; a fraction of that existing in the original population and of that induced by crossing the population to the inbred. If the disequilibrium (D) in the original population is positive (i.e., changes in gametic frequencies are in the same direction as that induced by crossing to the inbred) then total gametic disequilibrium in the F_1, and the genotypic contrasts in the F_2 and backcross generations, will be maximized. If D is negative, then the F_1

disequilibrium and the magnitude of the marker genotypic contrasts will be reduced. Marker allele and QTL frequencies will also affect the magnitude of the disequilibrium and the magnitude of the genotypic contrasts. Whether the inbred is homozygous for T_1 or t_1, disequilibrium will be maximized when the parent population is homozygous for the marker and for QTL alleles not present in the inbred. This is the classic case of a cross between two inbreds.

How can these results be used to maximize the probability of identifying marker allele-QTL associations in population × inbred crosses? Two points are obvious. Because maximum linkage disequilibrium will be achieved when $q=w=1$ (for the case when the inbred is $M_1 M_1 t_1 t_1$) or when $p=x=1$ (for the case when the inbred is $M_1 M_1 T_1 T_1$) and when $r=0$, marker loci and a QTL with alleles for which the population lacks the allele in the inbred should be chosen. Sufficient numbers of markers should be used to ensure that markers will be located as close to any QTL alleles as possible. This should enhance the probability of finding markers tightly linked to a QTL.

The choice of type of progeny to use for identifying marker-QTL associations is conditioned not only be the theoretical magnitude of the contrast in a particular generation, but also by the error associated with the contrast. In general, errors will be higher for contrasts dependent on individual plant data such as F_2, per se, and backcross, per se, than for contrasts were replicated progenies can be grown. In general, expected values of per se contrasts are approximately twice as large as those of testcross contrasts. However, testcross progenies can be replicated whereas, except for recombinant inbreds and F_3 lines, per se progenies, cannot. Recombinant inbreds have the largest contrasts. However, they suffer the disadvantage of requiring several more generations to produce than the other types of progenies. In choosing the type of progeny to use to identify marker-QTL associations, the impact of each of these factors on the objectives of identifying marker-QTL associations needs to be considered.

Because populations often are not useful for direct isolation of inbreds by selfing, they are usually crossed to inbreds and desirable traits from the population combined with those of the inbred. In many cases it may be desirable to backcross once to the inbred prior to selfing to develop new lines (Dudley 1982). In such cases, use of backcross progenies for identification of marker-QTL associations would be advantageous.

Two other points need to be considered. Ideally, the frequency of the favorable QTL allele in the population needs to be high and if marker loci and the QTL are in linkage disequilibrium, the disequilibrium should be in the same direction as that induced by crossing the population to the inbred. Populations improved by recurrent selection should be ideal for such studies because the frequencies of favorable QTL alleles should be high. If, in addition, the population has been identified as having alleles useful for improving the inbred, as was BS11 for improving Mo17 (Dudley 1988), then the frequency of favorable QTL alleles useful for improving the target hybrid should be high. Those marker alleles showing associations with a QTL in the population × inbred cross are likely to be those for which either there is no linkage disequilibrium in the population or those in which the disequilibrium is in the same direction as that induced by crossing the population to an inbred.

Identification of a QTL where the favorable allele is in the inbred is also useful. Where such alleles are identified, the frequency of the unfavorable allele in the population is likely to be high. Thus, for a new line derived from the population × inbred cross to be an improvement over the inbred, selection will need to be for favorable alleles both at the loci for which favorable alleles in the population have been identified and for those at which favorable alleles in the inbred have been identified.

The fact that marker-QTL associations can be identified in population × inbred crosses should allow the use of marker-facilitated selection to extract useful alleles from populations and incorporate them into inbreds while retaining favorable QTL alleles which were present in the inbred. An indication of the potential of this approach is found in Zehr et al. (1992) who demonstrated significant marker allele-QTL associations in the F_2 from the cross of BS11 × Mo17.

Acknowledgements. This research was supported by the University of Illinois Agricultural Experimental Station. The assistance of J. Moreno-Gonzalez in identifying an error in the original calculations and suggesting a transformation which greatly simplified the algebra is gratefully acknowledged.

References

Dudley JW (1982) Theory for transfer of alleles. Crop Sci 22:631–637

Dudley JW (1988) Evaluation of maize populations as sources of favorable alleles. Crop Sci 28:486–491

Falconer DS (1989) Introduction to quantitative genetics. 3rd edn. Longman Group Ltd, London, UK

Li CC (1955) Population genetics. The University of Chicago Press, Chicago, Illinois

Robbins RB (1918) Application of mathematics to breeding populations II. Genetics 3:73–92

Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8:552–560

Stuber CW (1992) Biochemical/molecular markers in plant breeding. Plant Bred Rev 9:37–62

Zehr BE, Dudley JW, Chojecki J, Saghai Maroof M, Mowers RP (1992) Use of RFLP markers to search for alleles in a maize population for improvement of an elite hybrid. Theor Appl Genet 83:903–911