

# Detection of QTL $\times$ environment interaction in maize by a least squares interval mapping method

MIRELLA SARI-GORLA\*<sup>†</sup>, TADEUSZ CALINSKI<sup>‡</sup>, ZYGMUNT KACZMAREK<sup>§</sup> & PAWEŁ KRAJEWSKI<sup>§</sup>

<sup>†</sup>Department of Genetics and Microbiology, University of Milan, Via Celoria 26, 20133 Milan, Italy, <sup>‡</sup>Department of Mathematical and Statistical Methods, Agricultural University of Poznan, Poland and <sup>§</sup>Institute of Plant Genetics, Polish Academy of Sciences, Poznan, Poland

In order to detect the linkage disequilibrium existing between alleles at a marker locus and alleles of a linked quantitative trait locus (QTL), a least squares interval mapping approach using multiple regression on marker data has been developed. It allows inclusion in the model of the parameters describing the experimental and environmental situation, so that the QTL  $\times$  environment effects can be tested. The method can also be applied using any general statistical package to data for which the usual normal distribution assumption does not hold, and where the use of weighted approaches is therefore required. A method to cope with the frequent problem in biological experiments of missing data was also used. The analysis was performed on data concerning two components of maize pollen competitive ability, obtained from an experiment over 2 years. The method, in comparison with the traditional single marker approach, has been shown to be more powerful in detecting QTLs and more precise in determining their map position. The analysis has identified QTLs expressed across years, putative QTLs with major effects and QTLs accounting for genotype  $\times$  environment interaction.

**Keywords:** interval mapping, maize, pollen, QTL  $\times$  environment interaction, RFLP.

## Introduction

Strategies for detecting polygenes are based on the linkage disequilibrium existing, owing to their physical linkage, between alleles at the marker locus and alleles of the linked QTL, and a number of methods for identifying the association between them have been proposed. A first approach is based on the analysis of single markers one at a time, comparing the marker class phenotypic means by analysis of variance or analysis of regression (Soller & Beckmann, 1983; Edwards *et al.*, 1987; Weller *et al.*, 1988; Stuber *et al.*, 1992; Zehr *et al.*, 1992). In a second approach, two or more marker loci are considered simultaneously. For this, originally, the method of maximum likelihood was adopted, in a way that allows estimation of the phenotypic effect and

provides a test statistic for a putative QTL at any given genetic location between two flanking markers, the method being called interval mapping (Lander & Botstein, 1989). It is a powerful tool for the mapping of QTLs, but the underlying maximum likelihood method is relatively complex and requires a special computer program to be applicable in practice. An alternative method, based on multiple regression, called least squares interval mapping, has been developed by Haley & Knott (1992). It has the advantage of being applicable with any general statistical package, e.g. GENSTAT in particular (Genstat 5 Committee, 1993). Moreover, it can easily be extended to more elaborate models, such as those with two or more QTLs, with environmental effects. When compared with the maximum likelihood method, it has given very similar results (Haley & Knott 1992). Apart from some simulation studies, the least squares interval mapping method

\*Correspondence. E-Mail: sari@imiucca.csi.unimi.it

has been applied successfully to certain real multi-environment phenotype data from a double haploid population of barley (Hayes *et al.*, 1993). Further extension of the single QTL interval mapping method has been considered independently by Jansen (1993, 1994), Jansen & Stam (1994), Zeng (1993, 1994) and Rodolphe & Lefort (1993). The main aim of these extensions is to increase the efficiency of detecting and the accuracy of mapping of multiple QTLs by employing multiple QTL models. The suggested general methods are based on regression models involving a large number of parameters to be estimated. Therefore, their practical use is restricted by the population size.

As the phenotype of an individual is also determined by the interaction of its genotype with the environment, even although a substantial proportion of QTLs affecting a quantitative trait is expected to be expressed in different environments, some QTLs important in one environment may not be as important in determining the phenotype in another environment. The interactions of individual QTLs with environments have been studied by an ANOVA procedure (Paterson *et al.*, 1988; Guffy *et al.*, 1989; Stuber *et al.*, 1992; Koester *et al.*, 1994), by comparing the frequency of identification of significant marker-QTL associations in different environments (Paterson *et al.*, 1991; Stuber *et al.*, 1992; Bubeck *et al.*, 1993; Schön *et al.*, 1994) or by direct estimation of QTL location using the least squares interval mapping method (Hayes *et al.*, 1993) and by the MQM mapping method (Jansen *et al.*, 1995). In general, marker locus  $\times$  environment interaction effects were found to be low, mainly concerning the magnitude of the effects detected in the diverse environments.

The present study was based on data from a 2-year experiment, performed on a population of recombinant inbred lines in order to detect QTLs controlling pollen competitive ability in maize. Our aims were:

- 1 to compare the information given by different, but closely related, methodologies for identifying QTLs, i.e. the traditional single marker approach, a method based on a theoretical model that takes into account the flanking markers and a method based on the same model but using markers as a covariate;
- 2 to identify QTLs expressed across years, putative QTLs with major effects and QTLs accounting for genotype  $\times$  environment interaction;
- 3 to evaluate the efficiency of the model in coping with the problem of missing marker observations.

## The model

### Genetic model

Let us assume that for a recombinant inbred (RI) line the genotypes at a number of marker loci located on a chromosome of length  $D$  cM are determined. Consider a fixed position,  $p$ , on this chromosome ( $0 < p < D$ ), at which a putative QTL,  $Q$ , is located. Denote by  $A$  and  $B$  two markers flanking  $Q$  from the left and right sides, respectively, and by  $p_A$  and  $p_B$  their positions on the chromosome. Inverting the procedure applied by Burr & Burr (1991) to construct the linkage map that was used in our study, we transform positions on the chromosome into expected fractions of recombinant RI lines by the formula (Haldane & Waddington, 1931):

$$R_A = \frac{2r_A}{1+2r_A}, \quad R_B = \frac{2r_B}{1+2r_B},$$

where  $r_A = (p - p_A)/100$  and  $r_B = (p_B - p)/100$  denote recombination fractions *per meiosis* between  $A$  and  $Q$  and between  $Q$  and  $B$ , respectively. Thus, following the authors of the map, we assume linearity between the distance in cM and recombination fractions, and we accept the assumption that  $r_A + r_B = r$ , the recombination fraction between  $A$  and  $B$ , which corresponds to the 'no double crossover' or 'full interference' model.

Let the genotype, for any locus  $G$  on the chromosome, be denoted by  $g(G)$ . Possible values of  $g(A)$  are  $A_1A_1$  and  $A_2A_2$ , of  $g(B)$  are  $B_1B_1$  and  $B_2B_2$ , of  $g(Q)$  are  $Q_1Q_1$  and  $Q_2Q_2$ . Probabilities of the two possible genotypes at  $Q$ , conditional on the genotypes at  $A$  and  $B$ , are presented in Table 1 (cf. Knapp *et al.*, 1990), in which the standardization factor  $R$  is equal to  $R_A + R_B$ .

Assume that expected values of the trait,  $y$ , being observed are  $\mu + \alpha$  for the genotype  $Q_1Q_1$ , and  $\mu - \alpha$

**Table 1** Probabilities of the two possible genotypes for the QTL,  $Q$

$g(A)g(B)$	Probability of	
	$g(Q) = Q_1Q_1$	$g(Q) = Q_2Q_2$
$A_1A_1 B_1B_1$	1	0
$A_1A_1 B_2B_2$	$R_B/R$	$R_A/R$
$A_2A_2 B_1B_1$	$R_A/R$	$R_B/R$
$A_2A_2 B_2B_2$	0	1

In the table, the standardization factor  $R$  is equal to  $R_A + R_B$ .

for the genotype  $Q_2Q_2$ , where  $\mu$  denotes a general mean and  $\alpha$  the genetic additive effect. Then, expected values for four possible genotypes at  $A$  and  $B$  can be expressed as

$$e(y) = \mu + x(g(A), g(B), p) \alpha, \quad (1)$$

where

$$x(g(A), g(B), p) = \begin{cases} 1, & g(A) = A_1A_1, g(B) = B_1B_1 \\ (R_B - R_A)/R, & g(A) = A_1A_1, g(B) = B_2B_2 \\ (R_A - R_B)/R, & g(A) = A_2A_2, g(B) = B_1B_1 \\ -1, & g(A) = A_2A_2, g(B) = B_2B_2. \end{cases}$$

If  $p$  is equal to the position of a marker  $A$ , obviously the flanking markers are  $A$  and  $A$ , and the above formula gives:

$$x(g(A), g(A), p) \equiv x(g(A)) = \begin{cases} 1, & g(A) = A_1A_1 \\ -1, & g(A) = A_2A_2 \end{cases}$$

**Model for experimental data**

The genetic model (eqn 1) described above will now be used to construct the model of observations of the trait for  $I$  inbred lines, compared in an experiment conducted at  $J$  environments, with  $K$  complete blocks in each trial. Denoting by  $y_{ijk}$  the observation concerning the  $i$ th line ( $i = 1, \dots, I$ ) coming from the  $j$ th environment ( $j = 1, \dots, J$ ) and the  $k$ th block ( $k = 1, \dots, K$ ), and assuming for observations from a block design experiment a fixed model, write:

$$e(y_{ijk}) = \mu + \eta_j + \beta_{jk} + \gamma_{ij}. \quad (2)$$

This model is a linear model involving a general mean  $\mu$ , environmental effects  $\eta_j$ ,  $j = 1, \dots, J$ , block effects  $\beta_{jk}$  nested within the environments,  $j = 1, \dots, J$ ,  $k = 1, \dots, K$ , and genotype effects  $\gamma_{ij}$ , also nested within environments,  $i = 1, \dots, I$ ,  $j = 1, \dots, J$ . Except for the constraints guaranteeing their uniqueness, parameters  $\gamma_{ij}$  are not assumed to have any special structure, so that is a maximal model with respect to the genotypic variation. Separation of genotype main effects and genotype  $\times$  environment interaction effects can be achieved by taking a version of eqn (2) with  $\gamma_{ij} = \gamma_i^{(0)} + \gamma_{ij}^{(1)}$ . On the other hand, the model with  $\gamma_{ij} = 0$ , that is

$$e(y_{ijk}) = \mu + \eta_j + \beta_{jk}, \quad (3)$$

assumes no genotypic effects at all.

Several models, intermediate between eqn (2) and

eqn (3), can be constructed using some decomposition of the genotypic effects, as

$$\gamma_{ij} = \phi_{ij} + \delta_{ij}. \quad (4)$$

In this formula, the parameter  $\phi_{ij}$  describes genotypic variability attributable to the regression of the phenotypic trait on the genotypes in  $L$  putative QTLs located at positions  $p_1, \dots, p_L$  (possibly on different chromosomes), and has the form:

$$\phi_{ij} = \sum_{l=1}^L x(g_i(A_l), g_i(B_l), p_l) \alpha_j^{(p_l)}, \quad (5)$$

where  $g_i(A_l)$  and  $g_i(B_l)$  denote genotypes observed for the  $i$ th line at markers  $A_l$  and  $B_l$  flanking the position  $p_l$ , and  $\alpha_j^{(p_l)}$  denotes the additive genetic effect characteristic for this position at the  $j$ th environment. Thus,  $\delta_{ij}$  in eqn (4) represents the genotypic variation not explained by the QTLs included in the construction of  $\phi_{ij}$ . Different forms of  $\phi_{ij}$  are considered below. Generally speaking, searching for the QTL model consists of looking for  $\phi_{ij}$  such that the model

$$e(y_{ijk}) = \mu + \eta_j + \beta_{jk} + \phi_{ij} \quad (6)$$

is not far, in terms of the goodness of fit, from the model (eqn 2), which means that  $\delta_{ij}$  is negligible. Alternatively, the aim of the analysis can be described as looking for such terms constituting eqn (5), which gives the best improvement of the model in eqn (6) over the model in eqn (3).

Positions  $p_1, \dots, p_L$  considered in eqn (5) can coincide with the positions of markers, which corresponds to assuming QTLs at marker loci, or can be taken as the positions of putative QTLs assumed to exist in the intervals between markers. We consider the intervals one at a time so that the number of QTLs of the first kind (marker covariates) is equal to  $L - 1 = M$  and the number of QTLs of the second kind is equal to 1. Thus, eqn (5) can be written as

$$\phi_{ij} = \sum_{l=1}^M x(g_i(A_l)) \alpha_j^{(p_l)} + x(g_i(A), g_i(B), p) \alpha_j^{(p)}, \quad (7)$$

which, when inserted into eqn (6), leads to the linear model assuming QTL  $\times$  environment interaction (i.e. the possibility of environment-specific additive genetic effects) at  $M$  marker loci and one locus located at some position  $p$ .

In the following, two models, simpler than eqns (6) and (7), will also be used. These are the model constructed under the hypothesis

$$H: \alpha_1^{(p)} = \alpha_2^{(p)} = \dots = \alpha_I^{(p)} = \alpha^{(p)},$$



having the parameter  $\phi_{ij}$  of the form

$$\phi_{ij} = \sum_{l=1}^M x(g_i(A_l)) \alpha_j^{(p_l)} + x(g_i(A), g_i(B), p) \alpha^{(p)}, \quad (8)$$

and therefore stating that the genetic additive effect of a putative QTL at the position  $p$  is not modified by the environment, and the model constructed under

$$H: \alpha_j^{(p)} = \alpha_j^{(p')} = \dots = \alpha_j^{(p'')} = 0,$$

having

$$\phi_{ij} = \sum_{l=1}^M x(g_i(A_l)) \alpha_j^{(p_l)}, \quad (9)$$

which says that no genetic additive effects exist at  $p$ .

Finally, to complete the formulation of the statistical model of observations, we assume that for all  $i$ ,  $j$  and  $k$

$$y_{ijk} = e(y_{ijk}) + e_{ijk},$$

where  $e_{ijk}$  denotes uncorrelated normally distributed experimental errors, each with expected values equal to 0 and a corresponding variance  $\sigma_{ijk}^2$ . With these assumptions, the parameters of the model in eqn (2) and genetic additive parameters appearing in eqn (7) can be estimated using the weighted least squares (WLS) method. Also, the fit of any pair of nested models can be compared using the appropriate  $F$ -statistic, constructed using as the denominator the residual mean square obtainable from the model in eqn (2). Of particular interest are comparisons of the following pairs of models:

- 1 eqn (3) vs. eqn (2), to test the general hypothesis about the existence of genotypic variation;
- 2 eqn (3) vs. eqn (6) with  $\phi_{ij}$  as in eqn (9), for different values of  $M$  and different selections of markers, to find their subset absorbing an appreciable portion of genotypic variation;
- 3 eqn (6) with  $\phi_{ij}$  as in eqn (9) vs. eqn (6) with  $\phi_{ij}$  as in eqn (7), for different positions  $p$ , to find those positions of loci that considerably improve the model, thus locating QTLs;
- 4 eqn (6) with  $\phi_{ij}$  as in eqn (8) vs. eqn (6) with  $\phi_{ij}$  as in eqn (7), for positions  $p$  at which a QTL has been declared, to assess the dependence of its additive effect on the environment.

To avoid estimation of parameters in models that describe the genotypic variation poorly, one can control the goodness of fit of models selected in 2 and 3 by comparing them with the model in eqn (2). Stages 1, 2, and 3–4 constitute the analysis which,

together with its results and specific properties, is described in detail in the next section. At stage 2, a version of the forward selection procedure was used, whereas for stage 3–4 the analysis of relevant models for all chromosomes and for  $p = 0, 1, \dots, D$  was performed.

## Application

### Data description

The data were produced in an experiment carried out in 2 years, with two blocks, on a set of 44 RI lines, kindly supplied by B. Burr, obtained from inbreds T232 and CM37 (Sari-Gorla *et al.*, 1992, 1995).

Two pollen traits were evaluated: pollen germinability and early growth, and late pollen tube growth rate, up to fertilization. The pollen characters were measured *in vivo* by means of a mixed pollination technique: equal amounts of the pollen to be tested and the pollen of a standard genotype (W22), carrying a genetic marker for coloured aleurone, were mixed and used to pollinate a female plant. As, in maize, the style length varies according to the positions of the flowers on the ear, increasing from the apex to the base, relative pollen tube growth rate (PTGR) can be expressed as the increase in the proportion of uncoloured kernels on the resulting ear. Each ear was divided into four or five segments eight kernels long, according to their length, in which the proportion of uncoloured kernels was computed. The regression surfaces were fitted by the WLS method (Grizzle *et al.*, 1969).

The proportion of uncoloured kernels in the apical segment of the ear gives a measure of pollen germinability (in terms of time and rate) and of early tube growth, which will be referred to as pollen grain germination (PGG).

Because of some experimental failures, the 'lines  $\times$  years' experimental design was not orthogonal. Variances of observations were calculated for PGG (on the basis of their binomial distribution) and for PTGR (taking into account that this trait is evaluated as the coefficient of regression of the proportion of uncoloured kernels on the ear segment). Reciprocals of the variances were used as weights to perform the WLS analysis for the models described previously.

For all lines included in the experiment, genotypes at 183 RFLP loci, located on 10 chromosomes, were known (Burr *et al.*, 1988); the number of markers on one chromosome varied from 14 to 26. The positions of markers in cM were determined using the linkage map published by Burr & Burr (1991).

### Some practical solutions

**Heterozygosity** The model described in eqn (1) assumes that the material is homozygous. Residual heterozygosity, caused probably by some unintentional selection, was, however, observed at marker loci, and at a level much greater than the level expected for the generation of selfing beyond the  $F_2$  (six) (Burr & Burr, 1991). A simple extension of the genetic model specified in eqn (1), based on theoretical expectations would not account for observed heterozygote frequencies. Therefore, for this analysis all marker observations classified as heterozygous were taken as missing, even although this could lead to biased estimates of location and effect, because heterozygous marker loci may well indicate heterozygous QTLs.

**Missing marker observations** For 60 markers, the genotypes of some lines (from 1 to 6) were unknown (1.8 per cent of marker observations were missing). Owing to the treatment of heterozygotes described above, the effective number of missing values in the data set taken was even higher and amounted to 6.6 per cent.

An approach to the problem of missing marker data was described by Jansen & Stam (1994). Their model can be analysed using the WLS method, with weights calculated as conditional probabilities of genotypic states, given the phenotype and genotypes at other marker loci. Because our model involves application of WLS with weights based on variances of observations, this approach would require double weighting and was not used in order to avoid further complications.

The method of treating missing marker observations proposed by Martinez & Curnow (1994), in the context of models using neighbouring markers, is more relevant for our analysis. The use of their approach means taking as flanking markers  $A$  and  $B$ , for the position  $p$  under consideration and for the line affected by the missing value, the nearest markers for which data are not missing. Thus, for some positions, the values of  $x(g_i(A), g_i(B), p)$  in eqn (7) are not determined on the basis of the same flanking markers for all  $i = 1, \dots, I$ . The situation for which one of the flanking markers cannot be found (missing observations at the chromosome ends), or is too far ( $> 50$  cM), requires further considerations. For all such cases, the calculation of the value of the explanatory variable  $x$  in eqn (7) is based on just one flanking marker, according to the formula (Simpson, 1989):

$$x(g(A), g(B), p) = \begin{cases} 1 - 2R_B, & g(A) \text{ unknown, } g(B) = B_1B_1 \\ -1 + 2R_B, & g(A) \text{ unknown, } g(B) = B_2B_2 \\ 1 - 2R_A, & g(A) = A_1A_1, g(B) \text{ unknown} \\ -1 + 2R_A, & g(A) = A_2A_2, g(B) \text{ unknown.} \end{cases}$$

Thus, the line is excluded from calculations only if no flanking markers can be found which are closer to  $p$  than 50 cM. In our data set, this situation happened only on chromosome 1, for line no. 1, positions  $p = 0, 1, \dots, 6$ .

**Selection of markers** Note that, for models involving multiple markers, a method of correcting missing observations is also needed at the marker selection stage. For this purpose, the method described above is equally suitable and was properly extended. All missing observations of marker genotypes were corrected on the basis of homozygous 'flanking' markers. The selection of markers constituting eqn (9) was then accomplished using the forward selection procedure, starting with  $M = 0$ , and adding at each step the marker that gave the best improvement of the model in terms of the  $F$ -statistic. The multiple correlation of each candidate marker with the ones already in the model was checked, and candidates with high correlation were rejected in order to avoid collinearities (Weisberg, 1985, p. 211). The selection process was stopped before nonsignificant  $F$ -values were obtained, taking into account possible overestimation of the markers' parameters while using the forward selection procedure (Weisberg, 1985, p. 214).

The forward selection procedure is certainly not the best method of selecting explanatory variables in multiple regression. Better algorithms, especially the methods of selecting the best subset of given size, are, however, numerically prohibitive for the number of variables involved here.

**Handling of selected markers** The mapping algorithm described earlier was enforced by the temporary elimination of markers included in the model and flanking the interval of current interest, as advocated by Jansen (1994) and Zeng (1994). It allows the role of markers to be re-examined, in the sense that the possibility of no marker  $\times$  environment interaction is studied in the same way as for all other positions. In a way, it corrects the overparameterization of the model, possible with the algorithm used at the marker selection stage.

**Controlling error rates** The first stage of mapping consists of testing the hypothesis of no QTL effect for all positions on the chromosome. We treat all these hypotheses as a family and test them simultaneously by the sequentially rejective Bonferroni procedure as described by Hochberg & Tamhane (1987, p. 57). This procedure allows control of the family-wise Type I error rate in the sense that it assures the probability of declaring no false QTLs to be at least  $1 - \varepsilon$ , with a small  $\varepsilon$  chosen in advance. In our calculations, the value of  $\varepsilon = 0.2$  was used, as a smaller value would make the family-wise Type II error rate too large.

At the second stage, positions of QTLs are checked for the presence of interaction with years. Here, far fewer hypotheses are tested on the chromosome, so the problem of controlling the Type I error rate for the family of no-interaction hypotheses is not so important. Additionally, for a fixed position, the two hypotheses tested on the first and the second stage form a family closed under intersection, and so, when testing at the level  $\varepsilon = 0.01$ , say, the probability of making no Type I error will be at least  $1 - \varepsilon$  (Hochberg & Tamhane, 1987, p. 54).

## Results

Full analysis of the data set was performed according to the methodology given earlier. At the first stage, the model in eqn (2) (with separation of genotypic main and interaction effects) was analysed for both traits (see Table 2). The presence of the genotype  $\times$  year interaction was established on the basis of the  $F$ -test.

The second stage consisted of the selection of markers that should be included in the model in eqn

(9). Results of the selection are presented in Table 3. The final model comprises regression on 10 markers for PGG and on six markers for PTGR. For both the traits, its fit is significantly better than that with the model with no markers, but not completely satisfactory when compared with the maximal model. Note that the percentage of genotypic variation explained by markers [measured by the ratio of the sum of squares for markers in eqn (9) and the sum of squares for genotypes in eqn (2)] is high.

Finally, the mapping of all 10 chromosomes was performed. For each position  $p = 0, 1, \dots, D$ , the following values were calculated.

- $F_1$ : value of the  $F$ -test statistic for the comparison of the model involving a QTL interacting with the years vs. the no-QTL model, as indicated in eqn (3) for the experimental data, with the corresponding  $P$ -value for that statistic.
- $F_2$ : value of the  $F$ -test statistic for the comparison of the model of a QTL  $\times$  year interaction vs. the model of no interaction for that QTL, as indicated in eqn (4), with the corresponding  $P$ -value.
- Estimates of environment-specific genetic additive parameters  $\alpha_1$  (first year) and  $\alpha_2$  (second year) in the model in eqn (7).
- Estimate of the common, for 2 years, genetic additive parameter  $\alpha$  in eqn (8).

Detailed results of mapping for PGG for chromosome 2 are presented graphically in Fig. 1. For the moment, suppose that the hypotheses are tested separately (not simultaneously). When testing at level 0.01, one interval, from 6 to 9 cM, can be declared as containing a QTL showing significant modification of its effect by the environment. This is indicated by significant values of both  $F_1$  and  $F_2$  (Fig. 1b, c). The estimate of  $\alpha$  in the first year is

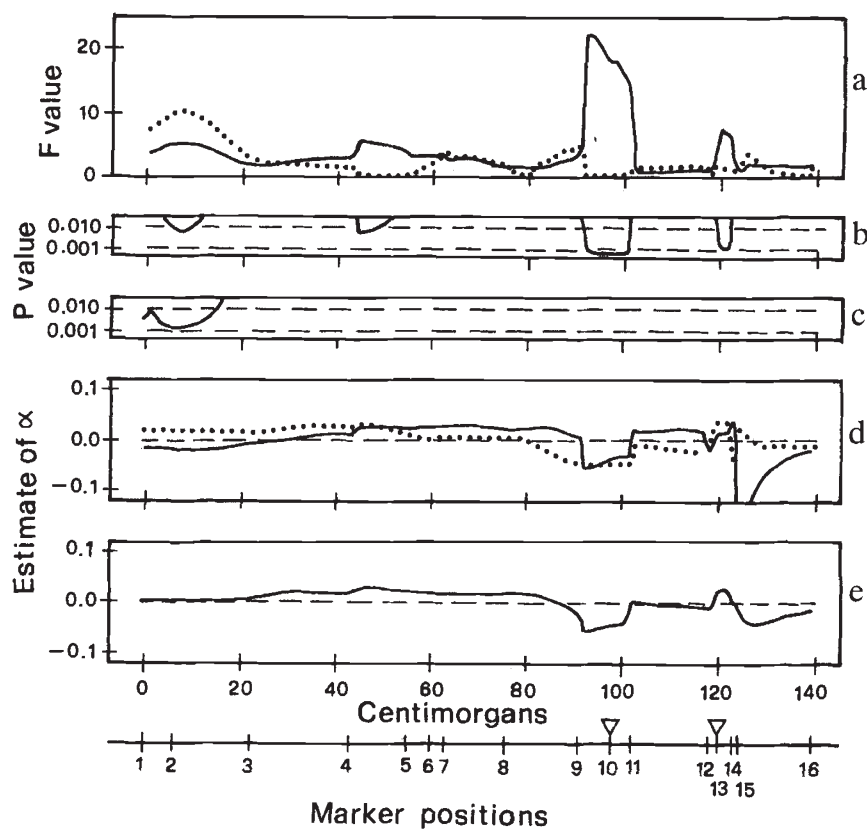
**Table 2** Analysis of variance for the maximal model of genotypic variation

Source of variation	Degrees of freedom	Mean square for	
		PGG	PTGR
Years	1	7429.20***	46.63**
Blocks within years	2	1648.43***	120.73***
Genotypes	43	246.92***	43.30***
Genotype $\times$ year interaction	38	39.08**	10.56**
Residual	71	17.34	5.44
Percentage of variance explained by the model		88	71

\*\*\* $P \leq 0.01$ , \*\* $P \leq 0.001$ .

**Table 3** Results for the model of regression on marker genotypes obtained by the forward selection procedure (markers enumerated in order of inclusion)

	Trait	
	PGG	PTGR
Markers included (marker no./chromosome no.)	2/8, 3/9, 16/1, 9/8, 13/2, 15/1, 3/12, 10/2, 13/10, 9/6	6/3, 3/7, 11/1, 3/4, 9/6, 3/9
<i>F</i> for comparison with the maximal model (equation 2)	2.19***	1.70*
Degrees of freedom	61, 71	69, 71
<i>F</i> for comparison with the model with no markers (equation 3)	28.33***	24.94***
Degrees of freedom	20, 71	12, 71
Percentage of variation explained by the model	83	62
Percentage of genotypic variation explained by markers	81	72

\* $P \leq 0.05$ , \*\*\* $P \leq 0.001$ .**Fig. 1** Results of the analysis for pollen grain germination (PGG), chromosome 2 of maize. (a) Value of  $F_1$  (the  $F$ -statistic for the existence of a QTL, solid line) and  $F_2$  (the  $F$ -statistic for the QTL  $\times$  year interaction, dotted line). (b)  $P$ -value corresponding to  $F_1$ . (c)  $P$ -value corresponding to  $F_2$ . (d) Estimate of the genetic effect  $\alpha$ , in the first year (solid line) and in the second year (dotted line). (e) Estimate of the genetic effect  $\alpha$  common for 2 years. Diagram below shows position of markers on the chromosome; markers acting as covariates are indicated by  $\nabla$ .

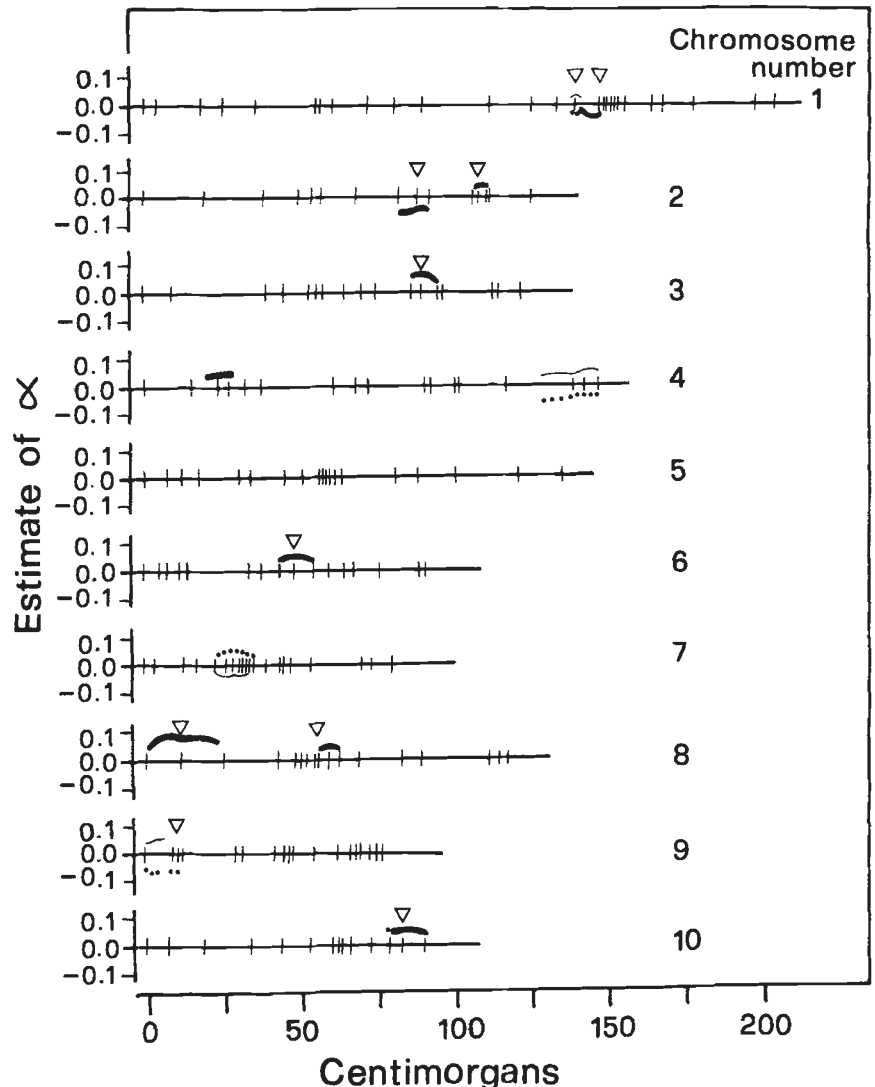


negative, and the estimate of  $\alpha$  in the second year is positive (Fig. 1d). Another interval, from 44 to 48 cM, can be reported as containing a QTL with the positive effect (Fig. 1e) common for 2 years, the conclusion being implied by significant values of  $F_1$  and nonsignificant values of  $F_2$ . Then two intervals appear that contain markers included in the model in the previous stage, namely the interval 92–102 cM, containing marker 10, and the interval 119–122 cM, containing marker 13. In both the intervals, the QTL  $\times$  year interaction is not present, which exemplifies the case of the re-examination of the marker's role in the model, as mentioned earlier.

One more interval shown in Fig. 1 requires special attention, despite the fact that the  $F$ -values in it are not significant. This is the interval 124–127 cM, in which untypically low estimates of  $\alpha$  in the first year

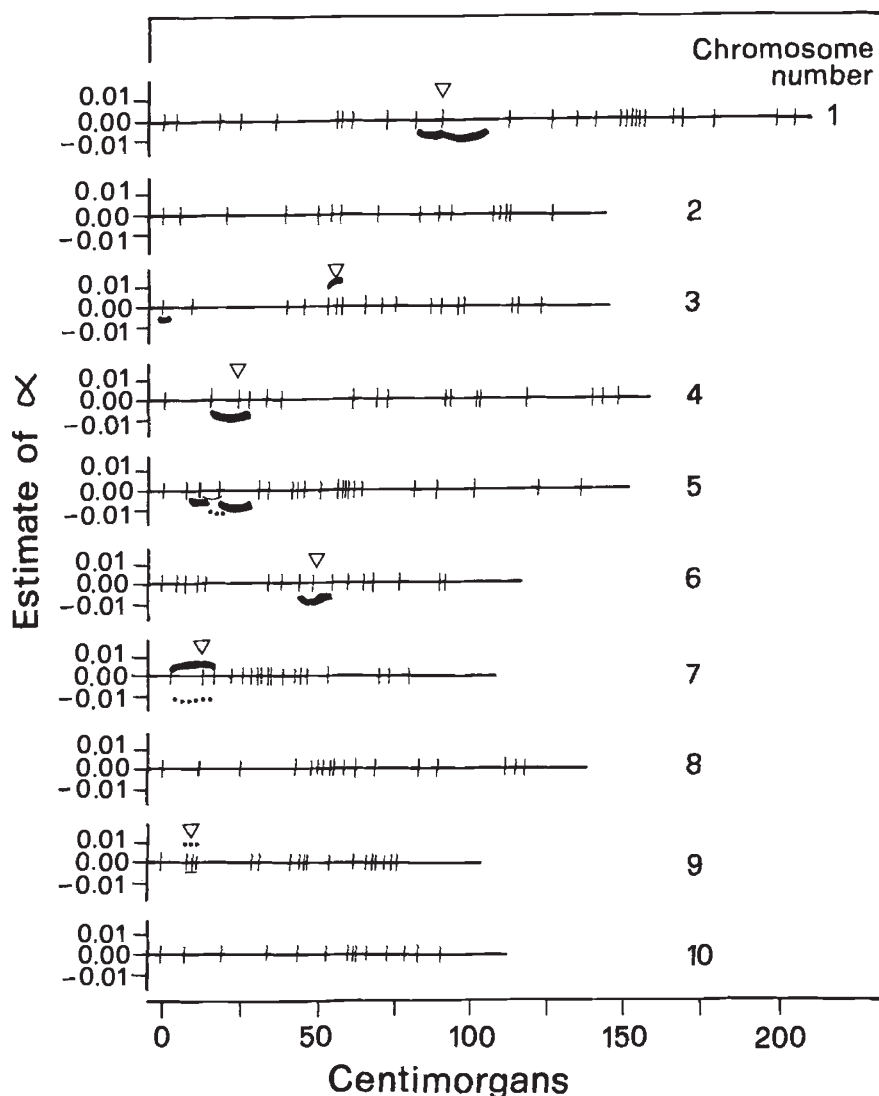
were obtained. This interval starts with the marker 15, which is located 4 cM from marker 13. Examination of the marker data showed substantial correlation ( $>0.9$ ) among markers 13, 14 and 15. It was concluded that the outlying estimates were obtained because of the failure of the WLS method to estimate properly parameters with two highly correlated explanatory variables in the model: marker 13, included in the model at the selection stage, and marker 15, serving as the flanking marker in the interval of interest. Such numerical problems did not occur in the other situations studied, and therefore it was not necessary to modify the analysis for this reason.

The amount of information obtained from the mapping performed is quite large and a way of summarizing it for practical purposes is needed.



**Fig. 2** Estimates of the genetic additive parameter  $\alpha$  – summary of the analysis undertaken for pollen grain germination (PGG), for 10 chromosomes of maize. Two lines indicate regions with significant QTL  $\times$  years interaction: the dotted line corresponds to estimates obtained for the first year and the thin line to estimates for the second year. The thick line indicates regions in which no interaction was detected but the common parameter  $\alpha$  was significant. Marks show markers' positions and  $\nabla$  indicates markers acting as covariates; the scale in cM is shown below.





**Fig. 3** Estimates of the genetic additive parameter  $\alpha$  – summary of the analysis undertaken for pollen tube growth rate (PTGR), for 10 chromosomes of maize. Two lines indicate regions with significant QTL  $\times$  years interaction: the dotted line corresponds to estimates obtained for the first year and the thin line corresponds to estimates for the second year. The thick line indicates regions in which no interaction was detected but the common parameter  $\alpha$  was significant. Marks show markers' positions and  $\nabla$  indicates markers acting as covariates; the scale in cM is shown below.

Figure 2 shows such a summary for PGG, for all 10 chromosomes, obtained on the basis of the simultaneous testing procedure described earlier.

In Fig. 2, the lines are drawn only within intervals in which the hypothesis of no QTL effect was rejected at the overall significance level of 0.2. In intervals in which significant (at the level 0.01) QTL  $\times$  year interaction was found, there are two lines, corresponding to estimates of  $\alpha$  in the two years. If interaction was not significant, just one line, corresponding to the estimate of  $\alpha$ , is shown. The graph allows regions suspected for the presence of QTLs, and regions in which the genetic effect of a putative QTL is modified by the environment (year), to be quickly identified. Owing to the testing procedure used, the probability of committing no error by declaring a false QTL on a given chromosome is at

least 0.8. Comparison of results obtained for chromosome 2 with those shown in Fig. 1 reveals that only the regions of markers 10 and 13 can be said to contain a QTL, if the family-wise Type I error rate is to be controlled.

A similar summary of results obtained for PTGR, for all 10 chromosomes, is shown in Fig. 3. For this trait, a modification of the mapping procedure, consisting of two runs, was tried. The first run was performed with the base model specified in Table 3, that is with the model involving marker  $\times$  year interaction parameters for all selected markers. The map obtained was checked, and significant interaction was found at the positions of marker 3 on chromosome 7 (3/7) and 3 on chromosome 9 (3/9). For the rest of the markers serving as covariates, the interaction was not significant. Thus, a new model with

main effects of markers 6/3, 11/1, 3/4 and 9/6 and interaction effects of markers 3/7 and 3/9 was constructed and used as the base model in the second run of mapping. In terms of the goodness of fit, this model is not appreciably different from the model shown in Table 3; it explains 70 per cent of the genotypic variation. The results of both mapping runs were virtually the same. So, for PTGR, very few intervals not containing markers acting as covariates were actually found.

## Discussion

Localization of QTLs for two components of competitive ability of maize pollen has been obtained from a 2-year experiment. The data had first been analysed using a one-at-a-time approach by means of regression analysis; only the loci that were significant but uncorrelated were included in a multiple regression model, in order to evaluate the cumulative contribution of the loci to the trait variability (Sari-Gorla *et al.*, 1992, 1995). In the present work, a different methodological approach was applied, in which the information derived from the flanking markers is taken into account, so that the estimates and test results for a putative QTL are obtained at any given genetic location between two flanking markers, providing a significance level profile for the presence of a QTL at any position on the chromosome covered by the markers.

The approach adopted here follows that of Haley & Knott (1992), with some extension to include in the model parameters reflecting the experimental and environmental situation. In that sense, the approach is similar to that of Hayes *et al.* (1993). In particular, the adopted model allows the QTL  $\times$  environment effects to be tested. Also, as far as the model construction and the mapping algorithm are concerned, the idea of using markers as covariates given by Jansen (1994) and Zeng (1994) was followed. The characters studied in the present work are intrinsically difficult to measure: this leads to data sets of variables for which the usual independence, constant variance and normal distribution assumptions do not hold. This requires sufficiently large samples on which the observed data (the proportions) are based, a suitable transformation of the data and the use of the weighted least squares method, with weights related to the unequal variances (Koch *et al.*, 1977). Therefore, it was desirable to apply a statistical technique that would account for the weights and would not require an additional system of weighting to be used, making the analysis too complex.

The statistical approach adopted, keeping the analysis relatively simple, at the same time indicates how various experimental and environmental situations can be taken into account in modelling the analysis. With regard to this, the present research can be considered as a step towards bringing the theoretical results on mapping of QTLs closer to plant breeding practice.

Comparisons between the results obtained by using a one-at-a-time analysis and interval mapping on the same set of data are reported in the literature (Paterson *et al.*, 1988; Stuber *et al.*, 1992; Bubeck *et al.*, 1993); they indicated that the two methods yielded virtually the same results; however, interval mapping was revealed to be more powerful. On the other hand, the methods involving markers as covariates are believed to be not sufficiently powerful when applied to limited experimental material (Zeng, 1994). For the present experiment, all three methods of analysis were used. The comparison of results reveals that, in general, similar chromosomal regions were detected; however, frequently the clusters of significant loci produced by the one-at-a-time analysis were resolved into two or more specific areas of significance.

Despite the limited experimental material and some imbalance of the data, the method adopted here has appeared to be sufficiently effective for revealing regions of chromosomes that might be of interest for marker-assisted selection.

Classical studies on quantitative traits measured  $G \times E$  interaction averaged across the entire genome; in recent studies, attempts have been made to discern the degree of  $G \times E$  interaction by individual QTLs by means of different approaches (Stuber *et al.*, 1992; Hayes *et al.*, 1993; Koester *et al.*, 1993; Schön *et al.*, 1994).

In the present study,  $G \times E$  interaction effects were detected for both the pollen characters. In some cases, the estimate of the genetic additive effects was negative in one of the years and positive in the other; in others, interaction was caused by different degrees of expression in the two years. Even although some QTL  $\times E$  effects could be identified by comparing the 2-year experiment results, others are not detectable in this way, as the single additive effects were not significant. Moreover, the statistical approach used allowed us to identify the more probable position on the chromosomes of the factor involved, whereas the bare comparison of the marker loci with different effects in the two years only indicates the presence of this type of effect in that chromosomal region.

Although data analysis and types of progeny can

provide improvements in the power of the experiment to detect QTLs, this is largely related to the number of progeny per marker class (Soller *et al.*, 1976; Beavis *et al.*, 1994). Thus, only a fraction of QTLs segregating for a given trait is expected to be identified in an experiment with a limited number of genotypes. However, our results, obtained from a small RI population, revealed an appreciable number of QTLs for both of the traits and in both the years, suggesting that the power to detect QTLs is also related to the presence of segregating progeny with high heritability for the character, as in the present case (Sari-Gorla *et al.*, 1992).

On the whole, the analysis allowed the identification and the localization on chromosomes of different categories of QTLs: those in which expression is modulated by the environmental effects and those that are expressed, albeit with variable expressivity, in both years, or that are independent of the environmental effects. The latter are probably 'major genes', QTLs on which the variation of the character mainly depends. This is important information in practical applications: the magnitude and the nature of QTL  $\times$  E interaction indicate the appropriateness of basing selection decisions on average QTL effects or environment-specific effects.

### Acknowledgements

The authors thank the two anonymous referees who reviewed the manuscript for their useful suggestions. This work has been supported by CNR (Consiglio Nazionale delle Ricerche), Italy, special project RAISA, paper no. 2799.

### References

- BEAVIS, W. D., SMITH, O. S., GRANT, D. AND FINCHER, R. 1994. Identification of quantitative trait loci using a small sample of topcrossed and  $F_4$  progeny from maize. *Crop Sci.*, **34**, 882–896.
- BUBECK, D. M., GOODMAN, M. M., BEAVIS, W. D. AND GRANT, D. 1993. Quantitative trait loci controlling resistance to grey leaf spot in maize. *Crop Sci.*, **33**, 838–847.
- BURR, B. AND BURR, F. A. 1991. Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations. *Trends Genet.*, **7**, 55–60.
- BURR, B., BURR, F. A., THOMPSON, K. H., ALBERTSON, M. C. AND STUBER, C. W. 1988. Gene mapping with recombinant inbreds in maize. *Genetics*, **118**, 519–526.
- EDWARDS, M. D., STUBER, C. W. AND WENDEL, J. F. 1987. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics*, **116**, 113–125.
- GENSTAT 5 COMMITTEE. 1993. *Genstat 5 Release 3 Reference Manual*. Clarendon Press, Oxford.
- GRIZZLE, J. E., STARMER, C. F. AND KOCH, G. G. 1969. Analysis of categorical data by linear models. *Biometrics*, **25**, 489–504.
- GUFFY, R. D., STUBER, C. W. AND EDWARDS, M. D. 1989. Dissecting and enhancing heterosis in corn using molecular markers. In: *Proc 25th Illinois Corn Breeders School*, pp. 99–120. Champaign, IL.
- HALDANE, J. B. S. AND WADDINGTON, C. H. 1931. Inbreeding and linkage. *Genetics*, **16**, 357–374.
- HALEY, C. S. AND KNOTT, S. A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.
- HAYES, P. M., LIU, B. H., KNAPP, S. J., CHEN, F., JONES, B., BLAKE, T. ET AL. 1993. Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm. *Theor. Appl. Genet.*, **87**, 392–401.
- HOCHBERG, Y. AND TAMHANE, A. C. 1987. *Multiple Comparison Procedures*. Wiley, New York.
- JANSEN, R. C. 1993. Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.
- JANSEN, R. C. 1994. Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, **138**, 871–881.
- JANSEN, R. C. AND STAM, P. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.
- JANSEN, R. C., VAN OOIJEN, J. W., STAM, P., LISTER, C. AND DEAN, C. 1995. Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. *Theor. Appl. Genet.*, **91**, 33–37.
- KNAPP, S. J., BRIDGES, W. C. AND BIRKED, D. 1990. Mapping quantitative trait loci using molecular marker linkage maps. *Theor. Appl. Genet.*, **79**, 583–592.
- KOCH, G. G., LANDIS, J. R., FREEMAN, J. L. AND FREEMAN, D. H., JR. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33**, 133–158.
- KOESTER, R. P., SISCO, P. H. AND STUBER, C. W. 1993. Identification of quantitative trait loci controlling days to flowering and plant height in two near isogenic lines of maize. *Crop Sci.*, **33**, 1209–1216.
- LANDER, E. S. AND BOTSTEIN, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.
- MARTINEZ, O. AND CURNOW, R. N. 1994. Missing markers when estimating quantitative trait loci using regression mapping. *Heredity*, **73**, 198–206.
- PATERSON, A. H., DAMON, S., HEWITT, J. D., ZAMIR, D., RABINOWITZ, H. D., LINCOLN, S. E. ET AL. 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics*, **127**, 181–197.
- 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 polymorphisms. *Nature*, **335**, 721–726.
- RODOLPHE, F. AND LEFORT, M. 1993. A multi-marker model for detecting chromosomal segments displaying QTL activity. *Genetics*, **134**, 1277–1288.
- SARI-GORLA, M., PÉ, M. E., MULCAHY, D. L. AND OTTAVIANO, E. 1992. Genetic dissection of pollen competitive ability in maize. *Heredity*, **69**, 423–430.
- SARI-GORLA, M., BINELLI, G., PÉ, M. E. AND VILLA, M. 1995. Detection of genetic factors controlling pollen–style interaction in maize. *Heredity*, **74**, 62–69.
- SCHÖN, C. C., MELCHINGER, A. E., BOPPENMAIER, J., BRUNKLAUS-JUNG, E., HERRMANN, R. G. AND SEITZER, J. F. 1994. RFLP mapping in maize: quantitative trait loci affecting testcross performance of elite European flint lines. *Crop Sci.*, **34**, 378–389.
- SIMPSON, S. P. 1989. Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theor. Appl. Genet.*, **77**, 815–819.
- SOLLER, M. AND BECKMAN, J. S. 1983. Genetic polymorphism in varietal identification and genetic improvement. *Theor. Appl. Genet.*, **67**, 25–33.
- SOLLER, M., GENIZI, A. AND BRODY, T. 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.
- STUBER, C. W., LINCOLN, S. E., WOLFF, D. W., HELENTJARIS, T. AND LANDER, E. S. 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics*, **132**, 823–839.
- WEISBERG, S. 1985. *Applied Linear Regression*, 2nd edn. Wiley, New York.
- WELLER, J. I., SOLLER, M. AND BRODY, T. 1988. Linkage analysis of quantitative traits in an interspecific cross of tomato (*Lycopersicon esculentum* × *Lycopersicon pimpinellifolium*) by means of genetic markers. *Genetics*, **118**, 329–339.
- ZEHR, I. E., DUDLEY, J. W., CHOJECKI, M. A., MAROOF, S. AND MOWERS, R. P. 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.
- ZENG, Z.-B. 1993. Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 10972–10976.
- ZENG, Z.-B. 1994. Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.