Theor Appl Genet (1992) 85:435–444

THEORETICAL AND APPLIED GENETICS

Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques

J. Moreno-Gonzalez

Centro de Investigaciones Agrarias de Mabegondo, Apartado 10, 15080 La Coruña, Spain

Received December 13, 1991; Accepted May 7, 1992
Communicated by A. R. Hallauer

Summary. The development of molecular markers has recently raised expectations for their application in selection programs. However, some questions related to quantitative trait loci (QTL) identification are still unanswered. The objectives of this paper are (1) to develop statistical genetic models for detecting and locating on the genome multi-QTL with additive, dominance and epistatic effects using multiple linear regression analysis in the backcross and F_n generations from the cross of two inbred lines; and (2) to discuss the bias caused by linked and unlinked QTL on the genetic estimates. Non-linear models were developed for different backcross and F_n generations when both epistasis and no epistasis were assumed. Generation analysis of marked progenies is suggested as a way of increasing the number of observations for the estimates without additional cost for molecular scoring. Some groups of progenies can be created in different generations from the same scored individuals. The non-linear models were transformed into approximate multivariate linear models to which combined stepwise and standard regression analysis could be applied. Expressions for the biases of the marker classes from linked QTL were obtained when no epistasis was assumed. When epistasis was assumed, these expressions increased in complexity, and the biases were caused by both linked and unlinked QTL.

Key words: Molecular markers – Epistatic effects – RFLP – Linkage – QTL

Introduction

Because classical methods of artificial selection have resulted in rapid genetic progress in plant and animal breeding for more than 100 years, most breeders rely on these methods and will presumably continue to do so until a new technology proves to be more efficient than the old. Futhermore, genetic gains in different crops (Fehr 1984) confirm that the theory underlying classical selection methods is still valid.

Recently, expectations for the application of molecular genetic markers to selection programs have risen. The theoretical basis for the identification of quantitative trait locus (QTL) effects associated with individual marker loci (Jayakar 1970; McMillan and Robertson 1974; Soller and Beckman 1983; Edwards et al. 1987; Cowen 1988) and flanking marker loci (Lander and Botstein 1989; Knapp et al. 1990) has been developed. However, examples of application of marker-assisted selection theory to crop selection programs are still scarce. Some questions related to QTL identification are still unanswered; e.g., (1) how can digenic epistatic interactions between single QTL be estimated? (2) how will linked QTL affect the estimation of gene effects? (3) how can different breeding strategies and statistical analyses of data improve the power of the genetic models? Some of these questions could be answered if an appropriate theory was developed. Some statistical techniques which have been developed to estimate QTL gene effects with markers are (a) the comparison of marker class means (Soller and Beckman 1983; Edwards et al. 1987; Cowen 1988), (b) the maximum likelihood estimation (Weller 1986; Lander and Botstein 1989; Luo and Kearsey 1989, 1991; Knapp et al. 1990) and (c) multiple linear regression (Cowen 1989; Knapp et al. 1990). Advantages and shortcomings of these methods have been discussed by Arus and Moreno-Gonzalez (1993). Since association between markers and QTL occurs mainly in linkage disequilibrium generations, the use of these generations is an advantage in these types of analysis.

The objectives of this paper are (1) to develop statistical genetic models for detecting and locating on the genome multi-QTL with additive, dominance and epistatic effects using multiple linear regression analysis in the backcross and F_n generations from the cross of two inbred lines, and (2) to discuss the bias caused by linked and unlinked QTL on the genetic estimates.

Genetic models

Definitions

Assume two inbred lines P_1 and P_2 with marker genotypes $M_1M_1 \ldots M_iM_i$ and $m_1m_1 \ldots m_im_i$ $(i=1,2,\ldots,n)$, respectively, and QTL genotypes $Q_1Q_1 \ldots Q_tQ_t$ and $q_1q_1 \ldots q_tq_t$ $(t \in T)$, respectively. Each pair of adjacent or flanking markers defines a marker chromosome segment S; e.g., markers M_i/m_i and M_{i+1}/m_{i+1} define segment S_i. The putative QTL lying in segment S_i is named Q_i/q_i. Not all marker segments carry a QTL. T is the set of subcripts of marker segments carrying a QTL. The F_1 cross between P_1 and P_2 has the following chromosome array:

$$\begin{array}{ccccccccc} & & M_i & & r_{1i} & Q_i & r_{2i} & & M_{i+1} \\ \_\_\_\_\_ & 0 & \_\_\_\_ & / & \_\_\_\_\_ & / & \_\_\_\_\_ & / & \_\_\_ \\ \_\_\_\_\_ & 0 & \_\_\_\_ & / & \_\_\_\_\_ & / & \_\_\_\_\_ & / & \_\_\_ \\ & & m_i & & & q_i & & & q_{i+1} \end{array}$$

where r_{1i} and r_{2i} are the recombination frequencies between loci M_i and Q_i and between Q_i and M_{i+1}, respectively. Because the chance of a double-crossover is at most 1% and 0.25% for 20 and 10 cM marker map distances, respectively, and much less if interference occurs as expected within small map distances (Strickberger 1985), only the no-double-crossover situation will be considered in this model (Knapp et al. 1990).

Let the QTL genotypes Q_iQ_i, Q_iq_i and q_iq_i be assigned the genotypic values $+a_i$, d_i and $-a_i$, respectively, where a and d stand for additive and dominance values, respectively (Falconer 1989); $r_i = r_{1i} + r_{2i}$, where r_i is the recombination frequency between M_i and M_{i+1}; $\rho_i = r_{1i}/r_i$ (Knapp et al. 1990).

Generations

The following generations were studied using this model.

1) Backcrossing of the F_1 cross (further, F_1 backcross generations B_1 and B_2), F_2 or advanced generations to both parents

If the F_1 cross, or random pollen from the F_2 or advanced (either random mating or selfing) generations is backcrossed, then individuals from the backcross generations can be scored for one of the eight marker classes in each marked segment (Table 1). Expected genetic values for the marker classes are shown in Table 1. When using pollen from the F_2 or advanced generations the recombination frequency between markers will increase relative to using pollen from the F_1 cross. Selfed families from scored individuals in the backcross generations can be used in replicated trials to reduce the environmental error component of the trait.

2) North Carolina Design III (NCIII) (Comstock and Robinson 1952)

Random individuals from the F_2 or advanced generations are scored and backcrossed to both parents. Nine marker classes for each segment can be distinguished

Table 1. Marker classes, expected frequency and expected genotypic values in the F_1 backcross generations B_1 and B_2 for a flanking marker model with no double-crossover

Backcross generation	Marker class	Coded class	Expected frequency[a]	Expected genotypic value[a]
B_1	$M_iM_iM_{i+1}M_{i+1}$	1	$\frac{1}{2}(1-r_i)$	a_i
	$M_iM_iM_{i+1}m_{i+1}$	2	$\frac{1}{2}r_i$	$(1-\rho_i)a_i+\rho_i d_i$
	$M_im_iM_{i+1}M_{i+1}$	3	$\frac{1}{2}r_i$	$\rho_i a_i+(1-\rho_i)d_i$
	$M_im_iM_{i+1}m_{i+1}$	4	$\frac{1}{2}(1-r_i)$	d_i
B_2	$M_im_iM_{i+1}m_{i+1}$	5	$\frac{1}{2}(1-r_i)$	d_i
	$M_im_im_{i+1}m_{i+1}$	6	$\frac{1}{2}r_i$	$(1-\rho)d_i-\rho a_i$
	$m_im_iM_{i+1}m_{i+1}$	7	$\frac{1}{2}r_i$	$\rho_i d_i-(1-\rho_i)a_i$
	$m_im_im_{i+1}m_{i+1}$	8	$\frac{1}{2}(1-r_i)$	$-a_i$

[a] $\rho_i = r_{1i}/r_i$ where $r_i = r_{1i} + r_{2i}$; r_{1i}, r_{2i} and r_i are the recombination frequencies between M_i and Q_i, Q_i and M_{i+1} and M_i and M_{i+1}, respectively. If pollen from the F_2 or advancing generations is used, the recombination frequency between markers will increase relative to using pollen from the F_1 cross

Table 2. Marker classes, expected frequency and expected genotypic value in the F_2 population and expected genotypic values of backcrosses from a design for a flanking marker model with no double-crossover

F_2 population				Expected genotypic values of backcrosses from a NCIII design	
Marker class	Coded class	Expected frequency	Expected genotypic value[a]	B_1	B_2
$M_iM_iM_{i+1}M_{i+1}$	1	$\frac{1}{4}(1-r_i)^2$	a_i	a_i	d_i
$M_iM_iM_{i+1}m_{i+1}$	2	$\frac{1}{2}r_i(1-r_i)$	$(1-\rho_i)a_i+\rho_i d_i$	$(1-\frac{1}{2}\rho_i)a_i+\frac{1}{2}\rho_i d_i$	$-\frac{1}{2}\rho_i a_i+(1-\frac{1}{2}\rho_i)d_i$
$M_iM_im_{i+1}m_{i+1}$	3	$\frac{1}{4}r_i^2$	$(1-2\rho_i)a_i+2\rho_i(1-\rho_i)d_i$	$(1-\rho_i)a_i+\rho_i d_i$	$-\rho_i a_i+(1-\rho_i)d_i$
$M_im_iM_{i+1}M_{i+1}$	4	$\frac{1}{2}r_i(1-r_i)$	$\rho_i a_i+(1-\rho_i)d_i$	$\frac{1}{2}[(1+\rho_i)a_i+(1-\rho_i)d_i]$	$\frac{1}{2}[-(1-\rho_i)a_i+(1+\rho_i)d_i]$
$M_im_iM_{i+1}m_{i+1}$	5	$\frac{1}{2}-r_i+r_i^2$	$\left\{1-\frac{2\rho_i(1-\rho_i)r_i^2}{1-2r_i+2r_i^2}\right\}d_i$	$\frac{1}{2}(a_i+d_i)$	$\frac{1}{2}(-a_i+d_i)$
$M_im_im_{i+1}m_{i+1}$	6	$\frac{1}{2}r_i(1-r_i)$	$-\rho_i a_i+(1-\rho_i)d_i$	$\frac{1}{2}[(1-\rho_i)a_i+(1+\rho_i)d_i]$	$\frac{1}{2}[-(1+\rho_i)a_i+(1-\rho_i)d_i]$
$m_im_iM_{i+1}M_{i+1}$	7	$\frac{1}{4}r_i^2$	$-(1-2\rho_i)a_i+2\rho_i(1-\rho_i)d_i$	$\rho_i a_i+(1-\rho_i)d_i$	$-(1-\rho_i)a_i+\rho_i d_i$
$m_im_iM_{i+1}m_{i+1}$	8	$\frac{1}{2}r_i(1-r_i)$	$-(1-\rho_i)a_i+\rho_i d_i$	$\frac{1}{2}\rho_i a_i+(1-\frac{1}{2}\rho_i)d_i$	$-(1-\frac{1}{2}\rho_i)a_i+\frac{1}{2}\rho_i d_i$
$m_im_im_{i+1}m_{i+1}$	9	$\frac{1}{4}(1-r_i)^2$	$-a_i$	d_i	$-a_i$

[a] $\rho_i = r_{1i}/r_i$ where $r_i = r_{1i}+r_{2i}$; r_{1i}, r_{2i} and r_i are the recombination frequencies between M_i and Q_i, Q_i and M_{i+1} and M_i and M_{i+1}, respectively

in the scored individuals from the F_2 or advanced generations. The relative frequencies of F_2 genotypes and their expected genotypic values when backcrossed to both parents are shown in Table 2. Expected genetic values of the sum and difference of the two backcrosses can be easily obtained from the last two columns in Table 2. The backcross sum is more powerful for estimating additive values, while the backcross difference is more powerful for dominance values. Since replicated backcross families can be tested with this mating design, it will reduce the environmental error component.

3) F_n generations

Scored individuals from the F_2 or advanced generations can be used in the analysis. Selfed families from these scored individuals will reduce the error component.

Mathematical models

Using the notation of Falconer (1989) and Mather and Jinks (1971), similar models to those of Knapp et al. (1990) can be developed.

1) No epistasis is assumed

From Tables 1 and 2, the following mathematical model for individuals from the F_1 backcross generations and NCIII can be written:

$$p_{jk} = \mu_0 + z_k + \sum_{i=1}^{n}(a_i x_i + d_i y_i + \rho_i a_i u_i + \rho_i d_i v_i) + \varepsilon_{jk} \quad (1)$$

where p_{jk} is the phenotypic value of individual or family j ($j=1,2,\ldots f$) in the backcross generation k ($k=1$ or 2); μ_0 includes the contribution of non-segregating QTL genes along with the average mean of all possible homozygous genotype combinations from segregating QTL that are included in the model; z_k is a class variable that accounts for the average mean effect of all genotypes in the backcross generation k from segregating QTL that are not included in the model; it also may include the environmental effect of generation k when tested separately; a_i and d_i are the additive and dominance values of the QTL associated with the marker segment S_i ($i=1,2,\ldots n$); x_i, y_i, u_i and v_i are dummy variables associated with a_i, d_i, $\rho_i a_i$, and $p_i d_i$, respectively; values of x_i, y_i, u_i and v_i for the marker classes of the F_1 backcross generations and NCIII backcrosses are shown in Table 3; ε_{jk} is the residual effect, which includes the environmental error effect, associated with individual or family j in generation k.

From Table 2, the following model for individuals from the F_2 generation can be derived:

$$p_j = \mu_0 + \sum_{i=1}^{n}(a_i x_i + d_i y_i + \rho_i a_i u_i + \rho_i d_i v_i + \rho_i^2 d_i w_i) + \varepsilon_j \quad (2)$$

where p_j is the phenotypic value of individual j ($j=1,2,\ldots f$) in the F_2 generation; μ_0 includes the contribution of nonsegregating QTL genes, along with the mean of all possible homozygous genotype combinations from segregating QTL in the model and the average mean of all genotypes in the the F_2 generation

Table 3. Values of the dummy variables x_i, y_i, u_i and v_i in the flanking marker genetic model with no double-crossover and non epistasis of Eq. 1[a] for the marker classes of the backcross generations B_1 and B_2 from the F_1 cross and from NCIII

Generation	Coded marker class[b]	Backcross from the F_1 cross				Backcross from NCIII			
		x_i	y_i	u_i	v_i	x_i	y_i	u_i	v_i
B_1	1	1	0	0	0	1	0	0	0
	2	1	0	−1	1	1	0	$-\frac{1}{2}$	$\frac{1}{2}$
	3	0	1	1	−1	1	0	−1	1
	4	0	1	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$
	5					$\frac{1}{2}$	$\frac{1}{2}$	0	0
	6					$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$
	7					0	1	1	−1
	8					0	1	$\frac{1}{2}$	$-\frac{1}{2}$
	9					0	1	0	0
B_2	1					0	1	0	0
	2					0	1	$-\frac{1}{2}$	$-\frac{1}{2}$
	3					0	1	−1	−1
	4					$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
	5	0	1	0	0	$-\frac{1}{2}$	$\frac{1}{2}$	0	0
	6	0	1	−1	−1	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$
	7	−1	0	1	1	−1	0	1	1
	8	−1	0	0	0	−1	0	$\frac{1}{2}$	$\frac{1}{2}$
	9					−1	0	0	0

[a] $y_{jk} = \mu_0 + z_k + \Sigma(a_i x_i + \Sigma d_i y_i + \Sigma \rho_i a_i u_i + \Sigma \rho_i d_i v_i) + \varepsilon_{jk}$
[b] According to Tables 1 and 2

from QTL not in the model; w_i is a dummy variable associated with $\rho_i^2 d_i$; remaining variables and parameters have already been defined; values of x_i, y_i, u_i, v_i and w_i for the marker classes of the F_2 generation are shown in Table 4.

Values of variables y_i, v_i and w_i will be multiplied by $(1/2)^t$ in Eqs. 1 and 2 if the tested families were derived by selfing the initial individuals or families in the backcross or F_2 generations during t generations. If the number of marker segments is n, then the non-linear models of Eqs. (1) and (2) have a maximum of $3n+2$ and $3n+1$ parameters to be estimated, respectively.

Table 4. Values of the dummy variables x_i, y_i, u_i v_i and w_i in the flanking marker genetic model with no double-crossover and non epistasis of Eq. 2[a] for the marker classes of the F_2 generation

Coded marker class	x_i	y_i	u_i	v_i	w_i
1	1	0	0	0	0
2	1	0	−1	1	0
3	1	0	−2	2	−2
4	0	1	1	−1	0
5	0	1	0	$-b_i$[c]	b_i[c]
6	0	1	−1	−1	0
7	−1	0	2	2	−2
8	−1	0	1	1	0
9	−1	0	0	0	0

[a] $y_j = \mu_0 + \Sigma(a_i x_i + d_i y_i + \rho_i a_i u_i + \rho_i d_i v_i + \rho_i^2 d_i w_i) + \varepsilon_j$
[b] According to Tables 1 and 2
[c] $b_i = 2r_i^2/(1 - 2r_i + 2r_i^2)$; r_i is the recombination frequency between flanking markers in segment S_i

2) Epistasis

Digenic interactions among loci are assumed. The following equation can be written for the backcross generations:

$$p_{jk} = \mu_0 + z_k + \sum_{i=1}^{n} (a_i x_i + d_i y_i + \rho_i a_i u_i + \rho_i d_i v_i) + \sum_{i<m}\sum A_{im}(x_i + \rho_i u_i)(x_m + \rho_m u_m) + \sum_{i \neq m}\sum AD_{im}(x_i + \rho_i u_i)(y_m + \rho_m v_m) + \sum_{i<m}\sum D_{im}(y_i + \rho_i v_i)(y_m + \rho_m v_m) + \varepsilon_{jk} \quad (3)$$

where A_{im}, AD_{im} and D_{im} are the additive × additive, additive × dominance and the dominance × dominance gene interaction between loci i and m as defined by Mather and Jinks (1971), respectively.

For the F_2 generation, the equation will be:

$$p_j = \mu_0 + \sum_{i=1}^{n} (a_i x_i + d_i y_i + \rho_i a_i u_i + \rho_i d_i v_i + \rho_i^2 d_i w_i)$$

$$+\sum_{i<m}\sum A_{im}(x_i+\rho_i u_i)(x_m+\rho_m u_m)$$

$$+\sum_{i\neq m}\sum AD_{im}(x_i+\rho_i u_i)(y_m+\rho_m v_m+\rho_m^2 w_m)$$

$$+\sum_{i<m}\sum D_{im}(y_i+\rho_i v_i+\rho_i^2 w_i)\cdot$$

$$\cdot(y_m+\rho_m v_m+\rho_m^2 w_m)+\varepsilon_j. \quad (4)$$

Equations 3 and 4 will estimate a maximum of $n(n-1)/2$ additional parameters for the A's and the D's, respectively, and $n(n-1)$ paramaters for the AD's.

Marked progeny generation analysis

The generation mean analysis as suggested by Hayman (1958, 1960) and Mather and Jinks (1971) can estimte additive, dominance and epistatic effects. One major disadvantage of this method is that estimates of positive and negative single gene effects are pooled through the genome in such way that they may cancel each other. In fact, estimates of additive effects for maize grain yield as indicated by Moreno-Gonzalez and Dudley (1981) were much less than expected because of the inherent limitations of the method.

A different approach is suggested. Additive, dominance and digenic epistatic gene effects directly associated with single marker segments could be estimated by applying a generation analysis to progenies that are molecularly marked. Since molecular scoring is more expensive than testing, individuals are first scored for molecular markers, and then selfed or backcrossed progenies are derived from them for testing of the phenotypes. Thus, the number of observations increases without additional cost for molecular scoring. Some examples of groups of progenies that can be created in different generations from the same scored individuals are (1) the F_3 progenies, the two groups of backcross families in NCIII and the second and reciprocal backcross family from each NCIII backcross family that can all be derived from original F_2 individuals; (2) the selfed progenies and the second and reciprocal backcross families derived from individuals in the F_1 backcross generations B_1 and B_2.

The following complete model can be written for fitting the data collected from testing marked progenies in different generations.

$$p_{jk}=\mu_0+z_k+\sum_i(a_i x_i+d_i y_i+\rho_i a_i u_i+\rho_i d_i v_i+\rho_i^2 d_i w_i)$$

$$+\sum_{i<m}\sum A_{im}(x_i+\rho_i u_i)(x_m+\rho_m u_m)$$

$$+\sum_{i\neq m}\sum AD_{im}(x_i+\rho_i u_i)(y_m+\rho_m v_m+\rho_m^2 w_m)$$

$$+\sum_{i<m}\sum D_{im}(y_i+\rho_i v_i+\rho_i^2 w_i)$$

$$\cdot(y_m+\rho_m v_m+\rho_m^2 w_m)+\varepsilon_{jk} \quad (5)$$

where the k subscript indicates the particular generation and the z_k term accounts for the mean of genotypes in generation k from segregating QTL not in the model, along with the environmental testing effect of generation k. Values of the w dummy variable are zero for the backcross generations. Values of the x, y, u and v dummy variables for the marker classes of the second and reciprocal backcrosses or any other generation can be easily obtained.

Statistical analysis

Some statistical computer program packages have now iterative subroutines that can fit non-linear models such as Eqs. 1–5 by an iterative least-square approach. They require specification of the fully expanded expression of the model, the names and starting values of the parameters and in some cases the first and second partial derivatives for each parameter in the model. If many parameters are included in the non-linear model, this kind of solution becomes extremely tedious and computer-time consuming.

To simplify the estimation procedure, an aproximate method is suggested. Make the following change of variables in the models:

$x_i'=x_i+\rho_i u_i$ for Eqs. 1–5;

$y_i'=y_i+\rho_i v_i$ for Eqs. 1, 2;

$y_i'=y_i+\rho_i v_i+\rho_i^2$ for Eqs. 3–5.

The parameter ρ_i is the probability that a recombinant gamete, say $M_i m_{i+1}$, carries the q_i allele. However, a particular recombinant gamete really carries either the Q_i or the q_i allele. Since individuals are fitted to the expected genotypic values of the recombinant marker classes, which depend on ρ_i, but not to their real genotypic values, then the parameter estimates will have an additional inherent error due to the recombinant marker classes even if ρ_i were known.

Because ρ is unknown and varies from 0 to 1, we assign an initial value of 0.5 to ρ_i for each i ($i=1,2,\ldots,n$). This initial asumption will affect the expected value of the recombinant marker classes 2, 3, 6 and 7 in the F_1 backcross generations and classes 2–4, 5 (not very much) and 6–8 in NCIII. Thus, they will bias the a and d estimates. The lower the recombination frequency, r_i, between flanking markers, the lower the bias will be, because a lower proportion of individuals affected by ρ_i will be involved in the estimates.

The non-linear models of Eqs. 1–5 will be changed to the following approximate multivariable linear

Table 5. Values of the dummy variables x' and y' for marker classes of backcross-derived families when $\rho = \frac{1}{2}$ and no epistasis is assumed

Coded marker class	Families					
	Selfing		Second backcross		Reciprocal backcross	
	x'	y'	x'	y'	x'	y'
1	1	0	1	0	0	1
2	0.5	0.25	0.75	0.25	−0.25	0.75
3	0.5	0.25	0.75	0.25	−0.25	0.75
4	0	0.5	0.5	0.5	−0.5	0.5
5	0	0.5	−0.5	0.5	0.5	0.5
6	−0.5	0.25	−0.75	0.25	0.25	0.75
7	−0.5	0.25	−0.75	0.25	0.25	0.75
8	−1	0	−1	0	0	1

model:

$$p_{jk} = \mu_0 + z_k + \sum(a_i x'_i + d_i y'_i) + \sum\sum A_{im} x'_i x'_m + \sum\sum AD_{im} x'_i y'_m + \sum\sum D_i y'_i y'_m + \varepsilon_{jk}. \quad (6)$$

The stepwise regression analysis (Draper and Smith 1981) available in some statistical packages can be performed on the model in Eq. 6. Cowen (1989) suggested stepwise multiple linear regression for analyzing RFLP-associated QTL data. The values of the dummy variables x' and y' when $\rho_i = 1/2$ are shown as an example in Table 5 for the selfing, second and reciprocal backcross families from the F_1 backcross generations. Likewise, appropriate values can be obtained for other sets of generations. If many parameters are to be estimated in the complete model, it will be advisable to leave out the interaction parameters *A*, *D* and *AD* and to perform a first analysis with the a and d parameters. On the basis of simulation studies, an α value of 0.005 seems to be a reasonable significance level to protect against high rates of false positives (experiment-wise type I errors) and large type II errors when 60 and 120 markers are studied.

Bias from linked and unlinked QTL

Several marked segments are usually available in the same chromosome for the analysis of the QTL effects. These segments likely do not segregate independently. Since the stepwise linear regression analysis incorporates sequentially, one by one, variables into the model, some of the questions to be asked are: (1) are the estimates of QTL effects which are included in the model biased by linked and unlinked QTL not yet in the model? (2) will a QTL already in the model prevent the entrance of a linked QTL not in the model? (3) are QTL effects influenced by other QTL when all are in the model? Attempts to answer these questions await further developments.

No epistasis

Four bias expressions affecting the marker classes of a linked QTL in the F_1 backcross generations were found (Appendix) when no epistasis was assumed:

$$G_i = \sum a_j(1 - 2R_{ij})(1 - 2\rho_j r_j)$$
$$G'_i = \sum a_j(1 - 2R_{ij})(1 - 2r_j + 2\rho_j r_j)$$
$$H_i = \sum d_j(1 - 2R_{ij})(1 - 2\rho_j r_j)$$
$$H'_i = \sum d_j(1 - 2R_{ij})(1 - 2r_j + 2\rho_j r_j) \quad (7)$$

where G_i and H_i are the biases for the marker classes of QTL *i*, which is in the model, due to the additive and dominance effects from those linked QTL that are not in the model and placed at one side (say, left-to-right DNA reading) of segment S_i, respectively; definitions of biases G'_i and H'_i are the same as G_i and H_i, respectively, but they are due to linked QTL *j* at the other side (say, reverse DNA reading) of S_i; subscript *j* refers to the linked QTL not included in the model; R_{ij} is the recombination frequency between the closest flanking markers of segments S_i and S_j.

These biases affect the estimates of gene effects and may prevent the entrance of linked QTL in the model when applying the stepwise linear regression analysis. Estimates of gene effects from linked QTL when first entering in the model are biased. They account for its own variation and part of that of linked QTL. The subsequent entrance of a linked QTL into the model would add little variation to the sum of squares already explained by the model. Therefore, it might be rejected. The biases can be included in the following model:

$$p_{jk} = \mu_0 + z_k + \sum_i (a_i x'_i + d_i y'_i + G_i g_i + G'_i g'_i + H_i h_i + H'_i h'_i) + \varepsilon_{jk} \quad (8)$$

Table 6. Values of the dummy variables associated to the biases of the marker classes of QTL *i*, that is included in the model, from linked QTL not included in the model for the F_1 backcross generations when no epistasis is assumed

Backcross	Coded marker classes	Dummy variables			
		g_i	g'_i	h_i	h'_i
B_1	1	1	1	−1	−1
	2	−1	1	1	−1
	3	1	−1	−1	1
	4	−1	−1	1	1
B_2	5	1	1	1	1
	6	−1	1	−1	1
	7	1	−1	1	−1
	8	−1	−1	−1	−1

where subscript i refers to QTL in the model; g_i, g'_i, h_i and h'_i are dummy variables associated to G_i, G'_i, H_i and H'_i, respectively. Their values for the marker classes of the F_1 backcrosses are shown in Table 6. Appropriate linear regression strategies would be able to estimate unbiased gene effects and to allow the entrance of linked QTL in the model.

Epistasis

Marker classes of QTL i, which is in the model, are biased with epistatic effects from both linked and unlinked QTL j, which are not included in the model. The following eight biases were found (Appendix):

$$A_i = \frac{1}{2}\sum_j A_{ij};$$

$$AD_i = \frac{1}{2}\sum_j AD_{ij};$$

$$DA_i = \frac{1}{2}\sum_j DA_{ij};$$

$$D_i = \frac{1}{2}\sum_j D_{ij};$$

$$A_i^L = \frac{1}{2}\sum_j A_{ij}(1-2R_{ij})(1-2r_j\rho_j);$$

$$AD_i^L = \frac{1}{2}\sum_j AD_{ij}(1-2R_{ij})(1-2r_j\rho_j);$$

$$DA_i^L = \frac{1}{2}\sum_j DA_{ij}(1-2R_{ij})(1-2r_j\rho_j);$$

$$D_i^L = \frac{1}{2}\sum_j D_{ij}(1-2R_{ij})(1-2r_j\rho_j);$$

where A_i, AD_i, DA_i and D_i are the pooled biases on QTL i from additive × additive, additive × dominance, dominance × additive and dominance × dominance gene interactions between locus i and remaining loci (linked and unlinked), which are not yet present in the model, respectively; A_i^L, AD_i^L, DA_i^L and D_i^L have similar definitions, but they are only caused by those linked loci not present in the model and placed at one side of marker segment S_i. Four additional biases, $A_i'^L$, $AD_i'^L$, $DA_i'^L$ and $D_i'^L$, caused by linked loci that are placed at the other side of the S_i marker segment, exist. They have, respectively, the same expressions than for A_i^L, AD_i^L, DA_i^L and D_i^L, but with ρ_j replaced by $(1-\rho_j)$.

The biases can be included in the following complete model:

$$p_{jk} = \mu_0 + z_k + \sum_i (a_i x'_i + d_i y'_i) + \sum_{i<}\sum_m A_{im}x'_i x'_m + \sum_{i\neq}\sum_m AD_{im}x'_i y'_m + \sum_{i<}\sum_m D_{im}y'_i y'_m + \sum_i (G_i g_i + G'_i g'_i + H_i h_i + H'_i h'_i) + \sum_i (A_i g_i + AD_i r_i + DA_i s_i + D_i t_i) + \sum_i (A_i^L q_i^L + AD_i^L r_i^L + DA_i^L s_i^L + D_i^L t_i^L + A_i'^L q_i'^L + AD_i'^L r_i'^L + DA_i'^L s_i'^L + D_i'^L t_i'^L) + \varepsilon_{jk} \quad (9)$$

where q_i, r_i, s_i, t_i, q_i^L, r_i^L, s_i^L, t'_i, $q_i'^L$, $r_i'^L$, $s_i'^L$ and $t_i'^L$ are dummy variables associated to A_i, AD_i, DA_i, D_i, A_i^L, AD_i^L, DA_i^L, D_i^L, $A_i'^L$, $AD_i'^L$ $DA_i'^L$ and $D_i'^L$, respectively. Their values for the marker classes of the F_1 backcross generations can be obtained from Table A2.

Discussion

The models proposed in this paper are similar to those developed by Knapp et al. (1990). However, some strategical modifications allow the proposed models to cope with different genetic situations: multiple QTL, epistasis, linkage and combined analysis of generations.

If no epistasis is assumed, Eq. 6 becomes

$$p_{ij} = \mu_0 + z_k + \sum_i (a_i x'_i + d_i y'_i). \quad (10)$$

Two and three parameters for each QTL are estimated in the proposed model [Eq. 10] and the Knapp et al. (1990) model, respectively. Thus, degrees of freedom of the residual sum squares are higher for Eq. 10 than for the Knapp et al. model when more than two QTL are estimated. This seems to suggest a greater power for the proposed model. However, additional computation for estimating the ρ_i is required in this model. The analysis of different generations (e.g., B_1 vs. B_2) will also allow a higher contrasting difference between parameters and a greater resolution power of the model.

The transformed linear models [Eqs. 6 and 10] will simplify the computational analysis for estimating genetic parameters. The transformation is based on assigning a value of 0.5 to ρ_i. Simulation studies from Moreno-Gonzalez (1993) illustrate how these models could be applied to datasets of individual plants and replicate selfed families using the SAS procedure STEPWISE (SAS 1985). The initial assignment ($\rho_i = 0.5$) did not cause a large shift in the residual mean square of the regression analysis nor a large bias in the parameter estimates (Moreno-Gonzalez 1993). However, the precision of the analysis would improve if estimates of true ρ_i values were obtained. A computer program for fitting data to the transformed models should be compiled using available regression subroutines. The following combination of stepwise and standard regression is suggested. When a significant QTL enters in the

model after performing a step in the stepwise regression, then standard regression analyses are performed in the new model by assigning values, in the range 0 to 1, to ρ_i. Values of ρ_i which have the smallest error mean square when fitting the model will be selected and retained for following steps. New values of the dummy variables for the recombinant marker classes will be computed using the ρ_i estimates. A final linear regression analysis then can be performed with a model that could include the interaction terms.

Epistasis and linked QTL increase the complexity of the models. Equation 9 is hard to manage since it includes too many parameters. Estimation of all of them may be impractical and meaningless. The researcher must make reasonable assumptions and decide which parameters should be retained and which ones left out in the model for each specific linear regression analysis.

Acknowledgements. This work was carried out during a stay at Iowa State University (ISU) while on leave from the Centro de Investigacions Agrarias de Mabegondo, La Coruna, Spain. It was sponsored in part by a fellowship from the Instituto Nacional de Investigaciones Agrarias (INIA), Spain and computer funds from the maize genetics laboratory at ISU. I am grateful to Drs. W. D. Beavis, J. W. Dudley and O. S. Smith for reading the manuscript.

Appendix

No epistasis

Let the QTL i be included in the model. The expected contribution of QTL j, that is not included in the model and placed at one side of segment S_i (say, left-to-right DNA reading), to the marker classes of linked QTL i in the F_1 backcross generations are derived from Table A1. Since the term $\frac{1}{2}(a_j+d_j)$ of the expected contribution is constant for marker classes 1–4 of backcross BC_1, it will be absorbed in the stepwise regression analysis by the parameter z_1 of Eq. 1. Similarly, the term $-\frac{1}{2}(a_j-d_j)$ of marker classes 5–8 of BC_2 will be absorbed by z_2. The expressions for a QTL j not included in the model and placed at the other side of segment S_i (say, reverse DNA reading) will be similar to the above expressions, but ρ_j changed by $1-\rho_j$. Then, four biases G_i, G_i', H_i and H_i', as defined in a previous section of the paper, were found.

Table A1. Expected contribution of QTL j, not included in the model, to the marker classes of linked QTL i, which is already in the model, for the F_1 backcross generations when no epistasis is assumed

Backcross	Coded marker class of QTL i[a]	QTL j not included in the model			Expected contribution of QTL j to the marker class of QTL i[b]
		Coded marker class	Conditional frequency on the marker class of QTL i	Expected genotypic value	
BC_1	1	1	$(1-R_{ij})(1-r_j)$	a_j	
		2	$(1-R_{ij})r_j$	$(1-\rho_j)a_j+\rho_j d_j$	
		3	$R_{ij}r_j$	$\rho_j a_j+(1-\sigma_j)d_j$	
		4	$R_{ij}(1-r_j)$	d_j	$\frac{1}{2}(a_j+d_j)+\frac{1}{2}(a_j-d_j)c$
	2	1	$R_{ij}(1-r_j)$	a_j	
		2	$R_{ij}r_j$	$(1-\rho_j)a_j+\rho_j d_j$	
		3	$(1-R_{ij})r_j$	$\rho_j a_j+(1-\rho_j)d_j$	
		4	$(1-R_{ij})(1-r_j)$	d_j	$\frac{1}{2}(a_j+d_j)-\frac{1}{2}(a_j-d_j)c$
	3				$\frac{1}{2}(a_j+d_j)+\frac{1}{2}(a_j-d_j)c$
	4				$\frac{1}{2}(a_j+d_j)-\frac{1}{2}(a_j-d_j)c$
BC_2	5				$-\frac{1}{2}(a_j-d_j)+\frac{1}{2}(a_j+d_j)c$
	6				$-\frac{1}{2}(a_j-d_j)-\frac{1}{2}(a_j+d_j)c$
	7	5	$(1-R_{ij})(1-r_j)$	d_j	
		6	$(1-R_{ij})r_j$	$(1-\rho_j)d_j-\rho_j a_j$	
		7	$R_{ij}r_j$	$\rho_j d_j-(1-\rho_j)a_j$	
		8	$R_{ij}(1-r_j)$	$-a_j$	$-\frac{1}{2}(a_j-d_j)+\frac{1}{2}(a_j+d_j)c$
	8	5	$R_{ij}(1-r_j)$	d_j	
		6	$R_{ij}r_j$	$(1-\rho_j)d_j-\rho_j a_j$	
		7	$(1-R_{ij})r_j$	$\rho_j d_j-(1-\rho_j)a_j$	
		8	$(1-R_{ij})(1-r_j)$	$-a_j$	$-\frac{1}{2}(a_j-d_j)-\frac{1}{2}(a_j+d_j)c$

[a] The derivation of the expected contribution of QTL j for marker classes 3, 4, 5 and 6 of QTL i is similar to that for classes 1, 2, 7 and 8

[b] $c=(1-2R_{ij})(1-2r_j\rho_i)$, where R_{ij} is the recombination frequency between the closest flanking markers of segments S_i and S_j

Table A2. Derivation of expected coefficients for the components of epistasis between QTL i, included in the model and QTL j, not included in the model, corresponding to the marker classes of QTL i, for the F_1 backcross generations

QTL i[a]		QTL j not included in the model			Expected coefficients of[b] epistatic components ($\times \frac{1}{2}$)			
Coded marker class	Expected genotype	Coded marker class	Conditional frequency on the marker class of QTL i	Expected genotype	A_{ij}	AD_{ij}	AD_{ji}	D_{ij}
1	Q_iQ_i	1	$(1-R_{ij})(1-r_j)$	Q_jQ_j				
		2	$(1-R_{ij})r_j$	$(1-\rho_j)Q_jQ_j+\rho_jQ_jq_j$				
		3	$R_{ij}r_j$	$\rho_jQ_jQ_j+(1-\rho_j)Q_jq_j$				
		4	$R_{ij}(1-r_j)$	Q_jq_j	$1+c$	$1-c$		
2	$(1-\rho_i)Q_iQ_i+\rho_iQ_iq_i$	1	$R_{ij}(1-r_j)$	Q_jQ_j				
		2	$R_{ij}r_j$	$(1-\rho_j)Q_jQ_j+\rho_jQ_jq_j$				
		3	$(1-R_{ij})r_j$	$\rho_jQ_jQ_j+(1-\rho_j)Q_jq_j$				
		4	$(1-R_{ij})(1-r_j)$	Q_jq_j	$(1-c)(1-\rho_i)$	$(1+c)(1-\rho_i)$	$(1-c)\rho_i$	$(1+c)\rho_i$
3	$\rho_iQ_iQ_i+(1-\rho_i)Q_iq_i$				$(1+c)\rho_i$	$(1-c)\rho_i$	$(1+c)(1-\rho_i)$	$(1-c)(1-\rho_i)$
4	Q_iq_i						$1-c$	$1+c$
5	Q_iq_i	5	$(1-R_{ij})(1-r_j)$	Q_jq_j				
		6	$(1-R_{ij})r_j$	$(1-\rho_j)Q_jq_j+\rho_jq_jq_j$				
		7	$R_{ij}r_j$	$\rho_jQ_jq_j+(1-\rho_j)q_jq_j$				
		8	$R_{ij}(1-r_j)$	q_jq_j			$-1+c$	$1+c$
6	$(1-\rho_i)Q_iq_i+\rho_iq_iq_i$	5	$R_{ij}(1-r_j)$	Q_jq_j				
		6	$R_{ij}r_j$	$(1-\rho_j)Q_jq_j+\rho_jq_jq_j$				
		7	$(1-R_{ij})r_j$	$\rho_iQ_jq_j+(1-\rho_j)q_jq_j$				
		8	$(1-R_{ij})(1-r_j)$	q_jq_j	$(1+c)\rho_i$	$(-1+c)\rho_i$	$(-1-c)(1-\rho_i)$	$(1-c)(1-\rho_i)$
7	$\rho_iQ_iq_i+(1-\rho_i)q_iq_i$				$(1-c)(1-\rho_i)$	$(-1-c)(1-\rho_i)$	$(-1+c)\rho_i$	$(1+c)\rho_i$
8	q_iq_i				$1+c$	$-1+c$		

[a] Derivation of expected contribution of QTL j for marker classes 3, 4, 7 and 8 of QTL i is similar to classes 1, 2, 5 and 6

[b] $c=(1-2R_{ij})(1-2r_j\rho_j)$, where R_{ij} is the recombination frequency between the closest flanking markers of segments S_i and S_j; A_{ij}, AD_{ij}, DA_{ij} and D_{ij} are the additive × additive, additive × dominance, dominance × additive and dominance × dominance epistasis between QTL i and j, respectively

Epistasis

The expected coefficients of the biases on the marker classes of QTL i, included in the model, caused by additive × additive (A_{ij}), additive × dominance (AD_{ij}), dominance × dominance (DA_{ij}) and dominance × dominance (D_{ij}) digenic interaction between QTL i and QTL j, not in the model, were derived in Table A2. Each interaction has two components. One component is caused by any kind of QTL j (linked or unlinked to QTL i); it does not involve the coefficient $c=(1-2R_{ij})\ (1-2r_jp_j)$, where R_{ij} is the recombination frequency between the closest flanking markers of segments S_i and S_j. The other additional component is only caused by a QTL j linked to QTL i; it does involve the coefficient c. For deriving the coefficients in Table A2, it was assumed that the linked QTL j was placed to one side of marker segment i (say, left-to-right DNA reading). If linked QTL j is placed to the other side of segment i (say, reverse DNA reading), the coefficients will be the same than in Table A2, but changing in the expression for c, ρ_j by $(1-\rho_j)$.

References

Arus P, Moreno-Gonzalez J (1993) Marker-assisted selection. In: Hayward MD Bosemark NO, Romagosa I (eds) Plant breeding principles and prospects. Chapman and Hall, London (in press)

Comstock RE, Robinson HF (1952) Estimation of average dominance of genes. In: Gowen JW (ed) Heterosis. Iowa State Colege Press, Ames, Iowa, pp 494–516

Cowen NM (1988) The use of replicated progenies in marker-based mapping of QTL's. Theor Appl Genet 75:857–862

Cowen NM (1989) Multiple linear regression analysis of RFLP data sets used in mapping QTLs. In: Helentjaris T, Burr B (eds) Development and application of molecular markers to problems in plant genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Draper NR, Smith H (1981) Applied regression analysis, 2nd edn. Wiley, New York

Edwards MD, Stuber CW, Wendel JF (1987) Molecular marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125

Falconer DS (1989) Introduction to quantitative genetics, 3rd edn. Longmans, London

Fehr WR (1984) Genetic contribution to yield gains of five major crop plants. CSSA Spec Publ 7. CSSA, Madison, Wis.

Hayman BI (1958) The separation of epistatic from additive and dominance variation in generation means. Heredity 12:371–390

Hayman BI (1960) The separation of epistatic from additive and dominance variation in generation means. II. Genetics 31: 133–146

Jayakar SD (1970) On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. Biometrics 26:451–464

Knapp SJ, Bridges WC Jr, Birkes D (1990) Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79:583–592

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

Luo ZW, Kearsey MJ (1989) Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. Heredity 63:401–408

Luo ZW, Kearsey MJ (1991) Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. II. Application to backcross and doubled haploid populations. Heredity 66:117–124

Mather K, Jinks JL (1971) Biometrical genetics. Chapman and Hall, London

McMillan I, Robertson A (1974) The power of methods for detection of major genes affecting quantitative characters. Heredity 32:349–356

Moreno-Gonzalez J (1993) Estimates of marker-associated QTL effects in Monte Carlo backcross generations using multiple regression. Theor Appl Genet 85:423–434

Moreno-Gonzalez J, Dudley JW (1981) Epistasis in related and unrelated maize hybrids determined by three methods. Crop Sci 21:644–651

SAS Institute (1985) SAS user' guide: statistics, basic version 5th edn. SAS Institute, Cary, N.C.

Soller M, Beckman JS (1983) Genetic polymorphism in varietal identification and genetic improvement. Theor Appl Genet 67:25–33

Strickberger MW (1985) Genetics, 3rd edn. Macmillan Publ, New York

Weller JI (1986) Maximum likelihood techniques for the mapping and analysis six quantitative trait loci with the aid of genetic markers. Biometrics 42:627–640