Theor Appl Genet (1992) 85:423–434

Estimates of marker-associated QTL effects in Monte Carlo backcross generations using multiple regression

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 decision of whether or not to use QTL-associated markers in breeding programs needs further information about the magnitude of the additive and dominance effects that can be estimated. The objectives of this paper are (1) to apply some of the Moreno-Gonzalez (1993) genetic models to backcross simulation data generated by the Monte Carlo method, and (2) to get simulation information about the number of testing progenies and mapping density in relation to the magnitude of gene effect estimates. Results of the Monte Carlo study show that the stepwise regression analysis was able to detect relatively small additive and dominance effects when the QTL are independently segregating. When testing selfed families derived from backcross individuals, dominance effects had a larger error standard deviation and were estimated at a lower frequency. Linked QTL require a higher marker mapping density on the genome and a larger number of progenies to detect small genetic effects. Reduction of the environmental error variance by evaluating selfed backcross families in replicate experiments increased the power of the test. Expressions of the number of progenies for detecting significant additive effects were developed for some genetic situations. The ratio of the within-backcross genetic variance to the square of a gene effect estimate is a function of the number of progenies, the heritability of the trait, the marker map density and the portion of the genetic variance explained by the model. Different values (from 0 to 1) assigned to ρ (relative position of the QTL in the marker segment) did not cause a large shift in the residual mean square of the model.

Key words: Genetic models – RFLP – Additive and dominance effects – Genetic linkage – Genetic simulation

Introduction

The wide use of molecular markers in crop selection programs will depend on their ability to increase the efficiency of selection. Major genes for disease resistance and other economical traits linked to isozyme markers have been identified. The theory for the identification of quantitative trait loci (QTL) underlying continuous variation in metric characters has been developed (Soller and Beckman 1983; Edwards et al. 1987; Lander and Botstein 1989; Knapp et al. 1990), while the theory on the integration of classical and marker-assisted selection into a selection index has been published by Lande and Thompson (1990). However, decisions on the use of QTL-associated markers in breeding programs requires information on the magnitude of additive and dominance effects that can be estimated in relation to genetic variance of the selection trait, number of progenies being evaluated and mapping marker density of the genome.

The Monte Carlo method is a valid sampling approach to simulate genetic situations that can test theory and provide useful information on genetic aspects insufficiently covered by theory. The objectives of this paper are (1) to apply some of the Moreno-Gonzalez (1993) genetic models to backcross simulation data generated by the Monte Carlo method, and (2) to get simulation information about the number of testing progenies and mapping density in relation to the magnitude of gene effect estimates.

Monte Carlo data and methods

Different sets of Monte Carlo data were generated to test some of the Moreno-Gonzalez (1993) models in the backcrosses of F_1 progeny to both parents. Either individual plant testing or testing of selfed families from the backcrosses were included. Sixty and 120 markers, each separated by a 0.2 and 0.11 recombination frequency, respectively, were simulated. Twelve chromosomes were involved in the simulations. Situations with both independent marker segments and linked segments saturating the entire genome were simulated. Random numbers from uniform distributions of the RANUNI SAS function (SAS 1985) were used to locate crossovers on chromosomes and produce gametes with random recombinant markers and QTL genotypes. The no-double-crossover situation (Knapp et al. 1990) was restricted to consecutive flanking markers, but no other restriction for number of crossovers on a chromosome was established. From 5 to 15 QTL were each located at a fixed position, ρ, within a marker segment. Independent and linked QTL were studied. Additive effects ranged from 1.5 to 0.2 arbitrary units for the favorable homozygous QTL. Negative values were assigned to the unfavorable homozygous genotypes. Each parental inbred included both favorable and unfavourable QTL. Either complete, partial or no dominance was arbitrarily assigned to the QTL. A random number from a normal distribution with mean zero and variance 1, N(0, 1), of the RANNOR SAS function (SAS 1985) was added to the genotypic value of each generated individual as an environmental error effect to simulate the phenotypic value in situations I–IV. Similarly, a random number from normal distributions, either N(0, 1.44) or N(0, 0.64), was added to the genotypic value of each individual in situation V.

All generated data were fitted by the SAS procedure PROC STEPWISE (SAS 1985) of multiple linear regression to the following simple model (Moreno-Gonzalez (1993)):

$$p_{jk} = \mu_0 + z_k + \sum_i (a_i x'_i + d_i y'_i) \tag{1}$$

where p_{jk} is the phenotypic value of individual or family $(j = 1, 2, \ldots f)$ in the backcross generation $k (k = 1$ or $2)$; μ_0 includes the contribution of nonsegregating 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 Q_i associated with the marker segment $S_i (i = 1, 2, \ldots n)$ (the S_i segments were consecutively numbered through the entire genome starting at one end of chromosome 1 until saturation, then chromosomes 2, 3, ... 12; not all S_i had located a corresponding Q_i); x' and y' are dummy variables [values of x' and y' were calculated for $\rho_i = \frac{1}{2}$ (Moreno-Gonzalez 1993)]. ρ_i indicates the relative position of the QTL in the marker segment S_i, $\rho_i = r_{1i}/r_i$ where r_{1i} and r_i are the recombination frequencies between the QTL i and the left-hand side flanking marker and between the two flanking markers, respectively. The constant value 10 was added to the phenotypic value of each individual in both backcrosses to simulate the contribution of the nonsegregating loci.

Results and discussion

Simulated data were generated for the following situations.

Situation I. Sixty marker segments with flanking markers separated by a 0.2 recombination frequency were assumed to segregate independently. QTL with additive and dominance effects were located arbitrarily within the segments (Table 1).

Estimates of additive effects larger than 0.75 were always significantly different from zero and estimated with a fair degree of unbiasedness when 500 plants were used (Table 1). Estimates of additive effects from 0.38 to 0.5 were significantly different from zero in more than 75% of the cases. Significant dominance effects were detected at high frequencies when testing individual plants, as in group 1 (Table 1). However, the frequency for small dominance effects decreased when selfed families from backcross individuals were tested, as in groups 2 and 3. In these cases, the error standard deviation associated with dominance effects was twice large as that associated with additive effects (Eqs. 3 and 4) because values of the dummy variable y' are half of the x' in the design matrix **X** (Appendix 1). Smaller error standard deviations for the dominance estimates would be expected if second and reciprocal backcrosses rather than selfed backcrosses were tested because of relative larger values of the dummy variable y' in those generations (Moreno-Gonzalez 1993, Table 5).

Situation II. Five consecutive marker segments on each of 12 chromosomes with flanking markers separated by a 0.2 recombination frequency were simulated. The values of recombination frequency and chromosome map distance do not correspond identically, however they become very close for small values. Thus, each chromosome was approximately 100 cM long. Five independent QTL with no dominance and no epistasis were each located on a chromosome (Table 2).

Data used for this situation had similar characteristics as those described by Lander and Botstein (1989). Stepwise linear regression as suggested in this paper was able to detect all QTL with a fair degree of unbiasedness when 250 progenies were used, 125 in each backcross. QTL with the largest effects were located at the simulated marker segment, while QTL with the smallest effects were located either at the simulated marker segment or at one of the two closest neighbor marker segments (Table 2). Half the estimates for Q_{16}, which had a 0.75 additive value and a true $\rho_{16} = 0.8$, were located at the closest segment, S_{17}. Of the estimates for Q_{22}, which had a 0.5 additive value and a true $\rho_{22} = 0.5$, 30% and 20% were located at segment S_{21} and S_{23}, respectively. Bias from linked QTL (Moreno-Gonzalez 1993) causes estimates of significant QTL effects at the closest neighbor marker segments and prevents entrance of the true QTL into the model.

This study seems to yield results similar to those obtained from the maximum likelihood method (ML) from Lander and Botstein (1989) for QTL with no dominance effects. Both methods detected QTL with large additive effects at the simulated marker segment. The stepwise linear regression detected smaller gene effects than the ML method at either the simulated marker segment or at the neighbor segments, likely because an F value with $\alpha = 0.005$ was used in the

Table 1. Estimates of additive and dominance gene effects and other results by stepwise linear regression analysis on a simulated genome with 60 independent marker segments and flanking markers separated by a 0.2 recombination value for groups of sets of backcross data generated by the Monte Carlo method with no epistasis and an F value of probability $\alpha < 0.005$

Group	Some characteristics of the data[a]	Simulated situation			Estimates and other results in the group[c]			
		QTL[b]	Effects		Significant estimates		Averaged estimates	
			Additive	Dominance	Additive	Dominance	Additive	Dominance
					%	%		
1	4 sets	Q_4	1.5	0.75	100	100	1.55	0.78
	500 plants/set	Q_8	−1.25	1.25	100	100	−1.24	1.25
	$\sigma_g^2 = 2.18$	Q_{12}	1.00	0.5	100	100	1.05	0.5
	$\sigma_E^2 = 1.00$	Q_{16}	−0.75	0.375	100	25	−0.65	0.35
	$h^2 = 0.45$	Q_{22}	0.5	0.5	100	100	0.56	0.51
	Individual plant testing	Q_{28}	−0.4	0.4	25	100	−0.36	0.39
2	10 sets	Q_4	1.5	0.75	100	100	1.51	0.85
	500 plants/set	Q_8	−1.25	1.25	100	100	−1.27	1.27
	$\sigma_g^2 = 1.63$	Q_{12}	1.00	0.5	100	30	0.98	0.75
	$\sigma_E^2 = 1.00$	Q_{16}	−0.75	0.375	100	0	−0.66	–
	$h^2 = 0.55$	Q_{22}	0.5	0.5	100	20	0.52	0.64
	Selfed backcross family testing	Q_{28}	−0.4	0.4	100	10	−0.44	0.52
3	3 sets	Q_1	0.5	0.25	67	0	0.67	–
	500 plants/set	Q_2	−0.5	0.25	100	0	−0.53	–
	$\sigma_g^2 = 1.67$	Q_3	0.5	0.25	67	33	0.54	0.66
	$\sigma_E^2 = 1.00$	Q_4	0.5	0.25	33	0	0.4	–
	$h^2 = 0.56$	Q_5	−0.5	0.5	67	67	−0.51	0.65
	Selfed backcross	Q_6	0.5	0.0	100	–	0.70	–
	family testing	Q_7	0.5	0.0	67	–	0.59	–
		Q_8	−1.25	1.25	100	100	−1.21	1.29
		Q_9	0.5	0.25	100	0	0.53	–
		Q_{10}	−0.5	0.0	67	–	−0.70	–
		Q_{12}	1.0	0.5	100	67	1.01	0.69
		Q_{16}	−0.75	0.375	100	0	−0.77	–
		Q_{22}	0.5	0.5	67	33	0.58	0.70
		Q_{28}	−0.4	0.4	67	0	−0.49	–
		Q_{38}	−0.38	0.0	33	–	−0.45	–

[a] σ_E^2 and σ_g^2 are environmental and pooled within-backcross genotypic variances, respectively

[b] ρ values were 0.7, 0.5, 0.2, 0.5, 0.5, 0.5, 0.9, 0.45, 0.5, 0.5, 0.35, 0.8, 0.5, 0.5 and 0.9 for Q_1, Q_2, Q_3, Q_5, Q_6, Q_7, Q_8, Q_9, Q_{10}, Q_{12}, Q_{16}, Q_{22}, Q_{28} and Q_{38}, respectively

[c] For groups 1, 2 and 3, respectively, percentages of false positives were 25%, 30% and 33% over sets and 0.46%, 0.56% and 0.72% over all possible false positives, and averaged residual mean squares (S_e^2) were 1.41, 1.23 and 1.45

regression and a $\log_{10}$ of the odds ratio (LOD) = 2.4 (Lander and Botstein 1989) equivalent to an F with a smaller α value was used in the ML method. However, the approach described herein is also able to detect QTL with dominance effects, while the ML method from Lander and Botstein (1989) is not.

Situation III. Five consecutive marker segments on each of 12 chromosomes with flanking markers separated by 0.2 recombination frequency were simulated. Each chromosome was approximately 100 cM long. Five linked QTL (Q_1–Q_5) in chromosome 1, 5 linked QTL (Q_6–Q_{10}) in chromosome 2 and 5 independent QTL in chromosomes 3, 4, 5, 6 and 8 were each located at a different marker segment (Table 3). Simulated QTL had both additive and dominance effects, but no epistatic effects.

The percentages of significant additive and dominance estimates were low for the linked QTL (Q_1–Q_{10}) when 500 progenies were used. A complex bias (Moreno-Gonzalez 1993) may affect the estimates. The averaged dominance estimates were overestimated. Two reasons may explain this: (1) since all simulated dominance effects were positive, the bias for any linked QTL entering the model will first be positive; (2) the dominance average was only computed with those estimates that were significant; these estimates should be high

Table 2. Estimates of gene effects and other results by stepwise linear regression analysis in a simulated genome with five linked marker segments on each of 12 chromosomes and flanking markers separated by 0.2 recombination frequency for a group of data sets generated by the Monte Carlo method from selfing the backcross generations with no dominance and an F value corresponding to a probability $\alpha < 0.005$

Group	Some characteristics of the data[a]	Simulated situation			Estimates and other results in the group[c]			
		QTL[b]	Effects		Significant estimates of simulated and close QTL[d]		Averaged estimates of simulated and close QTL	
			Additive	Dominance	Additive	Dominance	Additive	Dominance
					%			
1	10 sets	Q_4	1.5	0	100, 0	–	1.44	–
	250 plants/set	Q_8	1.25	0	100, 0	–	1.24	–
	$\sigma_g^2 = 1.41$	Q_{12}	1.00	0	90, 10	–	1.00	–
	$\sigma_E^2 = 1.00$	Q_{16}	0.75	0	50, 50	–	0.75	–
	$h^2 = 0.58$	Q_{22}	0.5	0	50, 50	–	0.56	–
	Independently segregating QTL, each in a different chromosome							

[a] σ_E^2 and σ_g^2 are environmental and within-backcross genotypic variances, respectively

[b] ρ values were 0.5, 0.45, 0.35, 0.8 and 0.5 for Q_4, Q_8, Q_{12}, Q_{16}, and Q_{22}, respectively

[c] Percentages of false positives were 20% over sets and 0.44% over all possible false positives, and averaged residual mean squares (S_e^2) was 1.30

[d] First and second numbers refer to percentages of significant QTL estimates at the right marker segment and the closest marker segment at any side of the simulated location, respectively

because a large error standard deviation is associated with them.

A high percentage of additive effects, even the small ones, from independent QTL (Q_{12}–Q_{38}) were estimated with 500 progenies. The percentages of significant dominance effects from independent QTL were low because of their high error standard deviation, as discussed above. Increasing the number of progenies increased the percentage of significant estimates and improved precision.

Situation IV. Ten consecutive marker segments on each of 12 chromosomes with flanking markers separated by a 0.11 recombination frequency were simulated. Each chromosome was approximately 110 cM long. 5 and 3 linked QTL (Q_1–Q_{10}) in chromosome 1, 5 and 3 linked QTL (Q_{12}–Q_{20}) in chromosome 2 and 5 and 10 independent QTL in the remaining chromosomes were each located at a different marker segment (Tables 4 and 5, respectively). Simulated QTL had both additive and dominance effects, but no epistatic effects.

A higher mapping density of linkage markers, as shown in Table 4, increased the percentage of significant estimates of the linked QTL in chromosomes 1 and 2 when compared to a lower mapping density (Table 3). Data sets with 1000 families, 500 for each backcross, and flanking markers separated by a 0.11 recombination value were able to detect significant small effects from QTL linked at about a 0.2 recombination frequency.

The percentage of significant additive and dominance estimates increased in loosely linked QTL (Table 5) compared to tightly linked QTL (Table 4). It seems that the higher the number of true linked QTL entering in the model, the lower the bias of the estimates.

Situation V. Ten consecutive marker segments on each of 12 chromosomes with flanking markers separated by a 0.11 recombination frequency were simulated. Each chromosome was approximately 110 cM long. Two loosely linked QTL (Q_1 and Q_{10}) in chromosome 1 and 11 unlinked QTL (Q_{16}–Q_{118}) in the remaining chromosomes were each located at a different marker segment (Table 6). Simulated QTL had both additive and dominance effect but no epistatic effects. Environmental variances (σ_E^2) with values 1.44, 1.44 and 0.64 were simulated in groups 1, 2 and 3, respectively. Reduction of the environmental variance increased the power of the test. A QTL with a 0.25 additive effect, which corresponds to 1/4.72 times the within-backcross genetic standard deviation, was significantly estimated in more than half of the experiments when the heritability of the trait was 0.59 (group 3).

Number of progenies

The detection of small genetic values depends on the number of progenies. The question – approximately how many progenies are needed to estimate a significant additive effect a_j at the $\alpha = 0.005$ level in at least

Table 3. Estimates of additive and dominance gene effects and other results by stepwise linear regression analysis in a simulated genome with five linked marker segments on each of 12 chromosomes and flanking markers separated by a 0.2 recombination frequency for groups of data sets generated by the Monte Carlo method with no epistasis and an F value of probability $\alpha < 0.005$

Group	Some characteristics of the data[a]	Simulated situation			Estimates and other results in the group[c]			
		QTL[b]	Effects		Significant estimates of simulated and close QTL[d]		Averaged estimates of simulated and close QTL	
			Additive	Dominance	Additive	Dominance	Additive	Dominance
					%			
1	10 sets	Q_1	0.5	0.25	0	0	–	–
	500 plants/set	Q_2	−0.5	0.25	20	20	−0.54	–
	$\sigma_g^2 = 1.67$	Q_3	0.5	0.25	70	50	0.57	1.12
	$\sigma_E^2 = 1.00$	Q_4	0.5	0.25	0	20	–	0.78
	$h^2 = 0.56$	Q_5	−0.5	0.5	10	20	−0.37	1.0
	Selfed backcross	Q_6	0.5	0.0	60	0	0.58	–
	family testing	Q_7	0.5	0.0	0	0	–	–
	Five linked QTL	Q_8	−1.25	1.25	50	100	−0.66	1.63
	Q_1–Q_5 in	Q_9	0.5	0.0	0	0	–	–
	chromosome 1, 5	Q_{10}	−0.5	0.25	0	20	–	1.02
	Q_6–Q_{10} in	Q_{12}	1.0	0.5	100	10, 30	0.97	0.67
	chromosome 2 and	Q_{16}	−0.75	0.375	80, 20	0	−0.76	–
	1 QTL in	Q_{22}	0.5	0.5	80	40	0.58	0.74
	chromosome	Q_{28}	−0.4	0.4	70	20	−0.55	0.68
	3, 4, 5, 6 and 8	Q_{38}	−0.38	0.0	40, 20	0	−0.38	–
2	5 sets	Q_1	0.5	0.25	60	20	0.24	0.66
	2000 plants/set	Q_2	−0.5	0.25	0	80	–	0.69
	$\sigma_g^2 = 1.67$	Q_3	0.5	0.25	100	0	0.58	–
	$\sigma_E^2 = 1.00$	Q_4	0.5	0.25	0	40	–	0.89
	$h^2 = 0.56$	Q_5	−0.5	0.5	60	60	−0.32	0.69
	Selfed backcross	Q_6	0.5	0.0	100	0	0.57	–
	family testing	Q_7	0.5	0.0	0	0	–	–
	Five linked QTL	Q_8	−1.25	1.25	100	100	−0.54	1.34
	Q_1–Q_5 in	Q_9	0.5	0.0	0	0	–	–
	chromosome 1, 5	Q_{10}	−0.5	0.25	60	40	−0.29	0.55
	Q_6–Q_{10} in	Q_{12}	1.0	0.5	100	40, 60	0.99	0.47
	chromosome 2 and	Q_{16}	−0.75	0.375	80, 20	0, 40	−0.78	0.40
	1 QTL in	Q_{22}	0.5	0.5	100	80, 20	0.60	0.46
	chromosome	Q_{28}	−0.4	0.4	100	80, 20	−0.43	0.44
	3, 4, 5, 6 and 8	Q_{38}	−0.38	0.0	80, 20	0	−0.36	–

[a] σ_E^2 and σ_g^2 are environmental and pooled within-backcross genotypic variances, respectively

[b] ρ values were 0.7, 0.5, 0.2, 0.5, 0.5, 0.5, 0.9, 0.45, 0.5, 0.5, 0.35, 0.8, 0.5, 0.5 and 0.9 for Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, Q_7, Q_8, Q_9, Q_{10}, Q_{12}, Q_{16}, Q_{22}, Q_{28}, and Q_{38}, respectively

[c] Percentages of false positives were 10% and 20% over sets and 0.25% and 0.50% over all possible false positives, and averaged residual mean squares (S_e^2) was 1.33 and 1.28 for groups 1 and 2, respectively

[d] First and second numbers refer to percentages of significant QTL estimates at the right marker segment and the closest marker segment at any side of the simulated location, respectively

half of the experiments – requires further developments.

The standard error associated with parameter estimates in multiple regression analysis (Draper and Smith 1981) is

$$s_i = \sqrt{S_e^2 c_{ii}} \tag{2}$$

where S_e^2 is the residual mean squares and c_{ii} is the diagonal term of the matrix $[\mathbf{X}'\mathbf{X}^{-1}]$; where $\mathbf{X}$ is the design matrix. The expected value of c_{ii} for independent QTL in a model where selfed backcross progenies are evaluated (Appendix 1) is:

$$c_{ii} = \frac{4}{N(1 - r_i)} \tag{3}$$

for additive effects and

$$c_{ii} = \frac{16}{N(1 - r_i)} \tag{4}$$

for dominance effects; where N is the number of observations and r_i is the recombination frequency

Table 4. Estimates of additive and dominance gene effects and other results by stepwise linear regression analysis in a simulated genome with ten linked marker segments on each of 12 chromosomes and flanking markers separated by a 0.11 recombination frequency for groups of data sets generated by the Monte Carlo method with no epistasis and an $F_{l,k,0.005}$ value

Group	Some characteristics of the data[a]	Simulated situation			Estimates and other results in the group[c]			
		QTL[b]	Effects		Significant estimates of simulated and close QTL[d]		Averaged estimates of simulated and close QTL	
			Additive	Dominance	Additive	Dominance	Additive	Dominance
					%			
1	10 sets	Q_1	0.5	0.25	30, 10	20	0.54	0.87
	500 plants/set	Q_4	−0.5	0.25	10	–	−0.53	–
	$\sigma_g^2 = 1.67$	Q_6	0.5	0.25	20, 50	30, 10	0.71	0.91
	$\sigma_E^2 = 1.00$	Q_8	0.5	0.25	0	10	–	1.22
	$h^2 = 0.56$	Q_{10}	−0.5	0.5	50	30, 20	−0.45	1.06
	Selfed backcross	Q_{11}	0.5	0.0	60, 30	0	0.58	–
	family testing	Q_{14}	0.5	0.0	20, 10	0	0.63	–
	Five linked QTL	Q_{16}	−1.25	1.25	90, 10	60, 40	−0.99	1.51
	Q_1–Q_{10} in	Q_{18}	0.5	0.0	0	0	–	–
	chromosome 1, 5	Q_{20}	−0.5	0.25	0	0	–	–
	Q_{11}–Q_{20} in	Q_{24}	1.0	0.5	90, 10	10, 0, 10	0.96	0.71
	chromosome 2	Q_{32}	−0.75	0.375	70, 10	10	−0.78	1.13
	and 1 QTL in	Q_{44}	0.5	0.5	60, 30	10, 0, 10	0.51	0.68
	chromosome	Q_{56}	−0.4	0.4	30, 30, 20	10, 10	−0.43	0.59
	3, 4, 5, 6 and 8	Q_{76}	−0.38	0.0	40, 30	0	−0.43	–
2	12 sets	Q_1	0.4	0.20	100	10	0.35	0.46
	1000 plants/set	Q_4	−0.4	0.20	17, 8	33, 25	−0.53	0.57
	$\sigma_g^2 = 1.57$	Q_6	0.5	0.25	50, 17	0	0.66	–
	$\sigma_E^2 = 1.00$	Q_8	0.5	0.25	42, 8	10	0.55	0.83
	$h^2 = 0.54$	Q_{10}	−0.5	0.5	83	42, 16	−0.47	0.72
	Selfed backcross	Q_{11}	0.5	0.0	58, 42	0	0.51	–
	family testing	Q_{14}	0.5	0.0	50, 25	0	0.55	–
	Five linked QTL	Q_{16}	−1.25	1.25	100	100	−1.17	1.39
	Q_1–Q_{10}	Q_{18}	0.5	0.0	50, 8	0	0.55	–
	in chromosome 1, 5	Q_{20}	−0.5	0.25	50	0	−0.55	–
	Q_{11}–Q_{20} in	Q_{24}	1.0	0.5	92, 8	67, 25	1.01	0.50
	chromosome 2	Q_{32}	−0.75	0.375	67, 33	25, 8	−0.87	0.41
	and 1 QTL in	Q_{44}	0.3	0.3	42, 42	0, 17	0.28	0.55
	chromosome	Q_{56}	−0.4	0.4	83, 17	8, 25, 8	−0.39	0.56
	3, 4, 5, 6 and 8	Q_{76}	−0.38	0.0	33, 67	0	−0.43	–
3	4 sets	Q_1	0.4	0.20	100	25	0.38	0.30
	2000 plants/set	Q_4	−0.4	0.20	75	0	−0.33	–
	$\sigma_g^2 = 1.57$	Q_6	0.5	0.25	75	50	0.52	0.53
	$\sigma_E^2 = 1.00$	Q_8	0.5	0.25	50, 50	25, 25	0.57	1.00
	$h^2 = 0.54$	Q_{10}	−0.5	0.5	100	75	−0.63	0.58
	Selfed backcross	Q_{11}	0.5	0.0	100	0	0.51	–
	family testing	Q_{14}	0.5	0.0	75, 25	0	0.50	–
	Five linked QTL	Q_{16}	−1.25	1.25	100	100	−1.20	1.40
	Q_1–Q_{10} in	Q_{18}	0.5	0.0	100	0	0.51	–
	chromosome 1, 5	Q_{20}	−0.5	0.25	75	0	−0.51	–
	Q_{11}–Q_{20} in	Q_{24}	1.0	0.5	100	100	1.02	0.51
	chromosome 2	Q_{32}	−0.75	0.375	100	50, 50	−0.76	0.48
	and 1 QTL in	Q_{44}	0.3	0.3	75, 25	75, 0, 25	0.29	0.31
	chromosome	Q_{56}	−0.4	0.4	100	100	−0.40	0.34
	3, 4, 5, 6 and 8	Q_{76}	−0.38	0.0	75, 0, 25	0	−0.39	–

[a] σ_E^2 and σ_g^2 are environmental and pooled within-backcross genotypic variances, respectively

[b] ρ values were 0.7, 0.5, 0.2, 0.5, 0.5, 0.5, 0.9, 0.45, 0.5, 0.5, 0.35, 0.8, 0.5, 0.5 and 0.9 for Q_1, Q_4, Q_6, Q_8, Q_{10}, Q_{11}, Q_{14}, Q_{16}, Q_{18}, Q_{20}, Q_{24}, Q_{32}, Q_{44}, Q_{56} and Q_{76}, respectively

[c] For groups 1, 2 and 3, respectively, percentages of additive false positives were 20%, 16% and 50% over sets and 0.26%, 0.22% and 0.66% over all possible positives; percentages of dominance false positives were 30%, 33% and 50% over sets and 0.39%, 0.44% and 0.66% over all possible positives

[d] First, second and third numbers refer to percentages of significant QTL estimates at the right marker segment and at the first and second closest marker segments at any side of the simulated location, respectively

Table 5. Estimates of additive and dominance gene effects and other results by stepwise linear regression analysis in a simulated genome with ten linked marker segments on each of 12 chromosomes and flanking markers separated by a 0.11 recombination frequency for groups of data sets generated by the Monte Carlo method with no epistasis and an $F_{t,k,0.005}$ value

Group	Some characteristics of the data[a]	Simulated situation			Estimates and other results in the group[c]			
		QTL[b]	Effects		Significant estimates of simulated and close QTL[d]		Averaged estimates of simulated and close QTL	
			Additive	Dominance	Additive	Dominance	Additive	Dominance
					%			
1	10 sets	Q_1	0.5	0.25	40, 50	0	0.58	–
	500 plants/set	Q_6	0.5	0.25	20, 30, 10	10, 0, 10	0.57	0.68
	$\sigma_g^2 = 1.67$	Q_{10}	−0.5	0.50	60	30, 30	−0.49	0.87
	$\sigma_E^2 = 1.00$	Q_{12}	0.5	0.0	40, 20, 10	0	0.42	–
	$h^2 = 0.56$	Q_{16}	−1.25	1.25	100	80, 20	−1.25	1.51
	Selfed backcross	Q_{24}	1.0	0.5	100	0, 30	1.03	0.65
	family testing	Q_{32}	−0.75	0.375	50, 50	0	−0.73	–
	Three linked QTL	Q_{44}	0.5	0.5	30, 50	10, 20	0.47	0.80
	Q_1, Q_6 and Q_{10} in	Q_{56}	−0.4	0.4	30, 60	10	−0.44	0.75
	chromosome 1, 2	Q_{64}	−0.5	0.25	60, 20, 20	0, 10	−0.50	0.66
	Q_{12} and Q_{16} in	Q_{76}	−0.38	0.0	60, 20	0	−0.43	–
	chromosome 2 and	Q_{88}	0.5	0.25	50, 30, 10	10	0.54	0.76
	1 QTL in remaining	Q_{94}	0.5	0.0	70, 20, 10	0	0.53	–
	chromosomes	Q_{110}	−0.5	0.0	80, 10	0	−0.50	–
		Q_{118}	0.5	0.25	70, 30	0, 20	0.56	0.68
2	8 sets	Q_1	−0.3	0.15	87	0	−0.31	–
	1000 plants/set	Q_6	0.5	0.25	75, 13	0, 13	0.51	0.65
	$\sigma_g^2 = 1.62$	Q_{10}	−0.5	0.50	87	38, 25	−0.48	0.58
	$\sigma_E^2 = 1.00$	Q_{12}	0.5	0.0	62, 13	0	0.46	–
	$h^2 = 0.55$	Q_{16}	−1.25	1.25	100	87, 13	−1.27	1.47
	Selfed backcross	Q_{24}	1.0	0.5	100	13, 25	1.01	0.57
	family testing	Q_{32}	−0.75	0.375	88, 13	0, 25	−0.74	0.48
	Three linked QTL	Q_{44}	0.5	0.5	50, 50	25, 37	0.47	0.54
	Q_1, Q_6 and Q_{10} in	Q_{56}	−0.4	0.4	63, 37	0, 25	−0.39	0.43
	chromosome 1, 2	Q_{64}	−0.5	0.25	87, 13	0, 13	−0.54	0.45
	Q_{12} and Q_{16} in	Q_{76}	−0.38	0.0	75, 25	0	−0.39	–
	chromosome 2 and	Q_{88}	0.5	0.25	75, 25	0, 25	0.50	0.55
	1 QTL in remaining	Q_{94}	0.5	0.0	50, 50	0	0.48	–
	chromosomes	Q_{110}	−0.5	0.0	87, 13	0	−0.46	–
		Q_{118}	0.5	0.25	87, 13	0	0.53	–

[a] σ_E^2 and σ_g^2 are environmental and pooled within-backcross genotypic variances, respectively

[b] ρ values were 0.7, 0.2, 0.5, 0.5, 0.45, 0.35, 0.8, 0.5, 0.5, 0.5, 0.9, 0.5, 0.9, 0.5 and 0.5 for Q_1, Q_6, Q_{10}, Q_{12}, Q_{16}, Q_{24}, Q_{32}, Q_{44}, Q_{56}, Q_{64}, Q_{76}, Q_{88}, Q_{94}, Q_{110} and Q_{118}, respectively

[c] For groups 1 and 2, respectively, percentages of additive false positives were 10% and 25% over sets and 0.20% and 0.49% over all possible positives; percentages of dominance false positives were 30% and 25% over sets and 0.59% and 0.49% over all possible positives

[d] First, second and third numbers refer to percentages of significant QTL estimates at the right marker segment and at the first and second closest marker segments at any side of the simulated location, respectively

between flanking markers. The a_i value has to be greater than $s_i t_{k,\alpha}(a_j > s_i t_{k,\alpha})$ to estimate significant effects different from zero at the 0.005 level, with 50% chance:

$$a_i > 2t_{k,\alpha}\sqrt{\frac{S_e^2}{N(1-r_i)}} \tag{5}$$

where $t_{k,\alpha}$ is the tabular t value for the k degrees of freedom of the residual mean squares and the chosen α significant level. Taking into account the above expressions and $t_{k,\alpha}^2 = F_{1,k,\alpha}$, then

$$N > \frac{4S_e^2 F_{1,k,\alpha}}{(1-r_i)a_i^2} \tag{6}$$

where $F_{1,k,\alpha}$ is the tabular F value for 1 and k degrees of freedom and an α significant value. In general, for a large number of degrees of freedom, estimates of significant additive effects at the α probability level in at least a β frequency of the experiments follows an

Table 6. Estimates of additive and dominance gene effects by stepwise regression in a simulated genome with ten linked marker segments on each of 12 chromosomes and flanking markers separated by a 0.11 recombination value for data sets generated by Monte Carlo with an $F_{l,k,0.005}$ value

Group	Some characteristics of the data[a]	Simulated situation			Estimates and other results in the group[c]			
		QTL[b]	Effects		Significant estimates of simulated and close QTL[d]		Averaged estimates of simulated and close QTL	
			Additive	Dominance	Additive	Dominance	Additive	Dominance
					%			
1	10 sets	Q_1	−0.3	0.15	40, 10	0	−0.41	–
	500 plants/set	Q_{10}	−0.5	0.50	50, 30, 10	40	−0.58	0.83
	$\sigma_g^2 = 1.49$	Q_{16}	−1.25	1.25	90, 10	30, 70	−1.23	1.29
	$\sigma_E^2 = 1.44$	Q_{24}	1.0	0.5	80, 10	20, 20	0.98	1.01
	$h^2 = 0.45$	Q_{32}	−0.75	0.375	50, 50	10	−0.71	0.78
	Selfed backcross	Q_{44}	0.5	0.5	60, 20	20, 20	0.57	0.74
	family testing	Q_{56}	−0.4	0.4	40, 50	10, 10	−0.45	0.71
	Two linked QTL	Q_{64}	−0.5	0.25	60, 20	0	−0.59	–
	Q_1 and Q_{10} in	Q_{76}	−0.38	0.0	10, 50	0	−0.44	–
	chromosome 1 and	Q_{88}	0.5	0.25	50, 40	0	0.49	–
	1 QTL in each of	Q_{94}	0.5	0.0	50, 40, 10	0	0.52	–
	remaining	Q_{110}	−0.5	0.0	80, 10, 10	0	−0.49	–
	chromosomes	Q_{118}	0.5	0.25	70, 10	0	0.49	–
2	10 sets	Q_1	−0.3	0.15	40, 40	10	−0.36	0.50
	1000 plants/set	Q_{10}	−0.5	0.50	60, 40	50, 10	−0.52	0.65
	$\sigma_g^2 = 1.49$	Q_{16}	−1.25	1.25	100	80, 10	−1.26	1.26
	$\sigma_E^2 = 1.44$	Q_{24}	1.0	0.5	100	0, 20, 20	1.00	0.63
	$h^2 = 0.45$	Q_{32}	−0.75	0.375	80, 20	20	−0.76	0.51
	Selfed backcross	Q_{44}	0.5	0.5	90, 10	20, 30	0.58	0.69
	family testing	Q_{56}	−0.4	0.4	40, 50	20, 10	−0.41	0.63
	Two linked QTL	Q_{64}	−0.5	0.25	100	0	−0.49	–
	Q_1 and Q_{10} in	Q_{76}	−0.38	0.0	50, 50	0	−0.37	–
	chromosome 1 and	Q_{88}	0.5	0.25	70, 30	0	0.53	–
	1 QTL in each of	Q_{94}	0.5	0.0	60, 40	0	0.45	–
	remaining	Q_{110}	−0.5	0.0	100	0	−0.45	–
	chromosomes	Q_{118}	0.5	0.25	60, 30	10, 0, 10	0.46	0.70
3	10 sets	Q_1	−0.3	0.15	60, 10	0	−0.40	–
	500 plants/set	Q_{10}	−0.5	0.50	90, 10	10, 30	−0.51	0.53
	$\sigma_g^2 = 1.39$	Q_{16}	−1.25	1.25	100	90, 10	−1.26	1.27
	$\sigma_E^2 = 0.64$	Q_{24}	1.0	0.5	90, 10	20, 30	0.97	0.57
	$h^2 = 0.59$	Q_{32}	−0.75	0.375	70, 30	10, 30	−0.72	0.62
	Selfed backcross	Q_{44}	0.5	0.5	70, 30	40, 30	0.38	0.72
	family testing	Q_{56}	−0.4	0.4	80, 20	20, 20	−0.42	0.60
	Two linked QTL	Q_{64}	−0.2	0.2	0, 10	0	−0.32	–
	Q_1 and Q_{10} in	Q_{76}	−0.38	0.0	20, 70, 10	0	−0.40	–
	chromosome 1 and	Q_{88}	0.5	0.25	80, 20	10, 10	0.48	0.55
	1 QTL in each of	Q_{94}	0.25	0.0	40, 30	0	0.33	–
	remaining	Q_{110}	−0.5	0.0	80, 10	0	−0.46	–
	chromosomes	Q_{118}	0.5	0.25	90, 10	0	0.44	–

[a] σ_E^2 and σ_g^2 are environmental and pooled within-backcross variances, respectively

[b] ρ values were 0.7, 0.5, 0.45, 0.35, 0.8, 0.5, 0.5, 0.5, 0.9, 0.5, 0.9, 0.5 and 0.5 for Q_1, Q_{10}, Q_{16}, Q_{24}, Q_{32}, Q_{44}, Q_{56}, Q_{64}, Q_{76}, Q_{88}, Q_{94}, Q_{110} and Q_{118}, respectively

[c] For groups 1, 2, and 3, respectively, percentages of additive false positives were 20%, 40% and 20% over sets and 0.32%, 0.64% and 0.32% over all possible positives; percentages of dominance false positives were 20%, 20% and 10% over sets and 0.27%, 0.27% and 0.14% over all possible positives; averaged residual mean squares (S_e^2) were 1.56, 1.60 and 0.78

[d] First, second and third numbers refer to percentages of significant QTL estimates at the right marker segment and at the first and second closest marker segments at any side of the simulated location, respectively

expression similar to that of Eq. 5 (Gill 1978):

$$a_i > 2(z_{1-\alpha/2} - z_\beta)\sqrt{\frac{S_e^2}{N(1-r_i)}} \qquad (7)$$

where z_β is the value of the variable of a standardized normal distribution such that the probability of a value lower than z_β is $1-\beta$.

Components of S_e^2 are:

$$S_e^2 = \sigma_E^2 + \sigma_{g'}^2 + \Phi^2 \qquad (8)$$

where σ_E^2 is the environmental error variance; $\sigma_{g'}^2$ is the part of the pooled within-backcross genetic variance σ_g^2 accounted for by QTL not yet included in the model, $\sigma_g^2 = \sigma_{g'}^2 + \sigma_{g''}^2$; where $\sigma_{g''}^2$ is the part of the genetic variance accounted for by QTL in the model; Φ^2 is a component due to the deviations of the assigned genotypic values from the real genotypic values for marker classes 2, 3, 6 and 7. Expected value of Φ^2 is $r(\sigma_g^2 - \sigma_{g'}^2) = r\sigma_{g''}^2$ (Appendix 2).

If all QTL were in the model, $\sigma_{g'}^2 = 0$ and the expected value of Φ^2 is $r\sigma_g^2$. Estimates of Φ^2, as $S_e^2 - \sigma_E^2$, in Tables 1, 2, 3 and 6 agreed fairly well with the expected value $r\sigma_g^2$.

Another important question is: approximately how many progenies, N_s, are required to detect the largest gene effects that account for a portion p of the genetic variance in a quantitative trait? Since not all effects are in the model, $\sigma_{g'}^2 = (1-p)\sigma_g^2$ and $\Phi^2 = rp_g^2$. The following expression is derived from Eqs. 6 and 8:

$$N_s > \frac{4F_{1,k,\alpha}\sigma_g^2\left[\dfrac{1}{h_b^2} - p(1-r)\right]}{(1-r)a_s^2} \qquad (9)$$

where σ_g^2 refers to the pooled within-backcross genetic variance; h_b^2 is the broad-sense heritability of the backcross generation and a_s is the lowest additive effect in the group of QTL that accounts for the portion p of the genetic variance.

ρ_i estimation

Gene effects in the above results were all estimated by assigning the value 0.5 to each ρ_i. Estimation of ρ_i for the QTL was carried out in one set of data from group 1 in Table 1, as follows: standard regression analyses were sequentially performed in a model that included only the significant variables after performing the stepwise regression. New x' values were computed by assigning to ρ_i values in the range from 0 to 1 for each significant QTL. The procedure started in the QTL with the largest effect (Q_4). Table 7 shows the residual mean squares from the analyses. The different ρ_i did not cause a large shift in the residual mean square, even for the QTL with the largest effect. This agrees with results of Knapp et al. (1990) who reported a stationary ridge in the response surface of SSE to different values of ρ_i. These results suggest that estimates of QTL effects are not great affected by the first assignment of 0.5 to each ρ_i. However a little improvement in the power of the model could be achieved if estimated values rather than the common 0.5 value were assigned to ρ_i.

Table 7. Residual mean squares from linear regression analysis in a model that included all significant QTL, for different values of ρ_i[a]

ρ_i[b]	Significant QTL				
	Q_4	Q_8	Q_{12}	Q_{16}	Q_{22}
0.1	1.449	1.410	1.290	1.312	1.300
0.2	1.410	1.382	$\underline{1.288}$	1.305	1.294
0.3	1.375	1.355	1.291	1.299	1.290
0.4	1.347	1.335	1.306	1.293	1.287
0.5	1.329	1.322	1.316	1.288	1.285
0.6	$\underline{1.322}$	$\underline{1.317}$	1.329	1.286	$\underline{1.284}$
0.7	1.326	1.319	1.344	$\underline{1.285}$	1.286
0.8	1.336	1.326	1.360	1.286	1.288
0.9	1.348	1.336	1.377	1.289	1.291

[a] The ρ_i values were sequentially assigned to each QTL from the largest (Q_4) to the lowest (Q_{22}) effect. The ρ_i with the lowest residual mean square (*underlined*) for each QTL was selected and retained for the following step of fitting

[b] The simulated ρ_i were 0.5, 0.45, 0.35, 0.8 and 0.5 for Q_4, Q_8, Q_{12}, Q_{16} and Q_{22}, respectively

General discussion

Unlinked QTL

The percentage of significant additive estimates with small effects for unlinked QTL (Tables 1–6) agreed fairly well with predictions derived from Eqs. 5 and 7. Standard deviations associated with estimates of additive and dominance effects in the simulation study (data not shown) almost exactly agreed with predictions derived from Eq. 2 taking into account Eqs. 3 and 4. Standard deviations of dominance effects were twice as large as those of additive effects, which explains the lower percentage of estimates with dominance effects and their upward bias.

Control of the residual mean square (S_e^2) in the model is essential to estimate lower genetic effects (Eqs. 5 and 7). The control can be made on the three components of Eq. 8. (1) The environmental error variance (σ_E^2), which also includes the genotype × environmental variance and the experimental error variance, is the most important component. Testing of selfed families in replicate experiments at several locations will reduce σ_E^2 and likely increase the heritability of the trait up to values similar to those simulated in this study. Heritabilities of family means between 0.4 and 0.82 for maize grain yield have been reported in selection experiments (Hallauer and Miranda Filho 1988). Table 6 shows the effect of reducing σ_E^2 from 1.44 in group 1 to 0.64 in group 3; this effect has been as important as increasing the number of testing families from 500 in group 1 to 1000 in group 2. (2) The Φ^2 component will be reduced by decreasing r; i.e., increasing the number of molecular markers on the genome. This will also have an additional effect on improving the power of the test by increasing the denominator in Eqs. 5 and 7. (3) No direct control can be made on the $\sigma_{g'}^2$ component, but it is reduced as more QTL enter into the model.

A quantitative trait is usually governed by many loci with small gene effects that generally are not equal and whose distribution is unknown. The stepwise regression incorporates gene effects sequentially into the model, starting with the largest effect. The number of progenies required for incorporating the largest gene effect into the model is derived from Eq. 9 when $p \approx 0$. An example: if $r = 0.11$ and $F_{l,k,0.005} \approx 8$ (for $k > 250$), Eq. 9 becomes:

$$N > \frac{36\sigma_g^2}{h_b^2 a_L^2} \tag{10}$$

where a_L is the largest gene effect and σ_g^2 is the pooled within-backcross genetic variance. Unfortunately, the ratio σ_g^2/a_L^2 is unknown before conducting the experiment; thus the number of progenies cannot be anticipated. Continuing with the same example, if $h_b^2 = 0.6$ and 500 selfed families are evaluated, then σ_g^2/a_L^2 has to be smaller than 8.33 to estimate the largest gene effect of the trait (Eq. 10). Since $\sigma_g^2 = \frac{1}{4}\Sigma(a_i^2 + \frac{1}{4}d_i^2)$, the condition $\Sigma(a_i^2 + \frac{1}{4}d_i^2) < 33.3\, a_L^2$ is required. This condition is expected to be met by many quantitative traits. In the unfavorable case where equal gene effects and complete dominance are assumed, traits controlled by less than 27 QTL will meet the condition.

Equation 9 can be used for computing the ratio σ_g^2/a_s^2 under different situations. An example: assume that gene effects with complete dominance follow the infinite and decreasing geometrical series (Lande and Thompson 1990) with ratio v, and the broad-sense heritability of the trait on family mean basis is 0.6. Which additional conditions have to meet the trait for estimating the highest gene effects that account for at least half of the genetic variance ($p = 0.5$) when 500 families are tested and $r = 0.11$? From Eq. 9 $\sigma_g^2/a_s^2 < 11.44$. Since $a_i = d_i$ and $a_L^2 = 2\ v^2 a_s^2$ in the geometrical series, then

$$\sigma_g^2 = \frac{5\Sigma a_i^2}{16} = \frac{5a_L^2}{16(1-v^2)} = \frac{10v^2 a_s^2}{16(1-v^2)} < 11.44a_s^2;$$

where $v < 0.9738$ is the ratio between a gene effect and the preceding. The condition for this ratio is likely met in some of the quantitative traits.

Therefore, according to the simulation study, the developed theory and the above examples, it is suggested that evaluation of 500 selfed backcross families, 250 for each backcross, in replicate trials with good control of the environmental error variance and flanking markers separated at about 10 cM could be used for estimating unlinked QTL in moderately complex traits by stepwise regression. Estimation of small gene effects in highly complex traits will require better and larger experiments.

Linked QTL

Linked QTL require more progenies than unlinked QTL to estimate significant effects as it can be seen by comparing Q_1–Q_{10} with the remaining QTL in Table 3 and Q_1–Q_{20} with the remaining QTL in Tables 4 and 5. An increase in the number of molecular markers on the genome will increase the percentage of significant estimates (compare Tables 3 and 4). As the linkage among QTL is reduced the percentage of significant estimates increases (compare Tables 4–6). The above results can be explained because the standard deviations associated with linked QTL estimates are larger than those with unlinked QTL (Eqs. 2–4). Since the residual mean square S_e^2 is also a factor of the error standard deviation (Eq. 2) for linked QTL, again reduction of the environmental error variance by evaluating replicate families in several environments and increase in the marker map density of the genome will raise the percentage of significant estimates for linked QTL.

False positives

True positives were considered if the significant QTL was estimated (1) either on the simulated segment or on the adjacent segment at any of both sides when a 0.2 recombination value is being studied, and (2) if the QTL was estimated on the simulated segment or on the two closest segments at any of both sides when a 0.11 recombination value is being studied. False positives were the remaining significant estimates corresponding to non-simulated QTL. Results of false positives (Tables 1–6) were expressed as an experimentwise type I error (percentage of the number of false positives over the number of sets in the group) and a comparisonwise type I error (percentage of the number of false positives over the number of all possible false positives in the genome). Results of the comparisonwise type I error in the simulation study agreed fairly well with the expected probability $\alpha = 0.5\%$, which was chosen for the F test. Likewise, results of the experimentwise type I error generally agreed with the expected value $1 - (1 - 0.005)^n$, where n is the number of possible false positives in the genome.

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 1

The following **X** design (data) matrix can be written for the model of Eq. 1:

Individual	Marker class	Expected frequency of marker class	Values of variables in the data matrix **X**								
			μ_0	z_k	...	x'_i	x'_{i+1}	...	y'_i	y'_{i+1}	...
1	1	$\frac{1}{2}(1-r)$	1	1	...	1	1	...	0	0	...
2	2	$\frac{1}{2}\mathrm{r}$	1	1	...	$\frac{1}{2}$	$\frac{1}{2}$	...	$\frac{1}{4}$	$\frac{1}{4}$	...
3	3	$\frac{1}{2}r$	1	1	...	$\frac{1}{2}$	$\frac{1}{2}$	...	$\frac{1}{4}$	$\frac{1}{4}$	...
4	4	$\frac{1}{2}(1-r)$	1	1	...	0	0	...	$\frac{1}{2}$	$\frac{1}{2}$	...
.	.	.	1	1	...	.	.	...	.	.	...
.	.	.	.	.	...	.	.	...	.	.	...
N_1	.	.	1	1	...	.	.	...	.	.	...
1	5	$\frac{1}{2}(1-r)$	1	-1	...	0	0	...	$\frac{1}{2}$	$\frac{1}{2}$	...
2	6	$\frac{1}{2}r$	1	-1	...	$-\frac{1}{2}$	$-\frac{1}{2}$	...	$\frac{1}{4}$	$\frac{1}{4}$	...
3	7	$\frac{1}{2}r$	1	-1	...	$-\frac{1}{2}$	$-\frac{1}{2}$	...	$\frac{1}{4}$	$\frac{1}{4}$	...
4	8	$\frac{1}{2}(1-\mathrm{r})$	1	-1	...	-1	-1	...	0	0	...
.	.	.	1	-1	...	.	.	...	.	.	...
.	.	.	.	.	...	.	.	...	.	.	...
N_2	.	.	1	-1	...	.	.	...	.	.	...

where N_1 and N_2 are the number of scored individuals in the backcross 1 and 2, respectively; μ_0 and z_k have already been defined in the section of Monte Carlo data and methods; x'_i, x'_{i+1}, ... and y'_i, y'_{i+1},... are dummy variables corresponding to the additive and dominance effects of QTL i, $i+1,\dots$, respectively; values of x' and y' are taken from Table 4 in Moreno-Gonzalez (1993); r is the recombination frequency between flanking markers. If the QTL are unlinked (independent) and the number of scored individuals in backcross 1 and 2 is the same, $N_1 = N_2 = N/2$, the following expectd **X'X** matrix (E[**X'X**]) is obtained:

$$E[\mathbf{X'X}] = \frac{N}{4}\begin{bmatrix} 4 & 0 & 0 & 0 & \dots & 1 & 1 & \dots \\ 0 & 4 & 2 & 2 & \dots & 0 & 0 & \dots \\ 0 & 2 & 2-r & 1 & \dots & 0 & 0 & \dots \\ 0 & 2 & 1 & 2-r & \dots & 0 & 0 & \dots \\ . & . & . & . & \dots & . & . & \dots \\ 1 & 0 & 0 & 0 & \dots & \frac{2-r}{4} & \frac{1}{4} & \dots \\ 1 & 0 & 0 & 0 & \dots & \frac{1}{4} & \frac{2-r}{4} & \dots \\ . & . & . & . & \dots & . & . & \dots \end{bmatrix}$$

where columns 1 and 2 correspond to μ_0 and z_k, respectively; columns 3 to $f+2$ correspond to variables x'_i associated with additive effects and columns $f+3$ to $f+h+2$ correspond to variables y'_i associated with dominance effects; f and h are the number of variables with additive and dominance effects in the model, respectively.

The expected inverse matrix was computed:

$$E([\mathbf{X'X}]^{-1}) = \frac{1}{N(1-r)}$$

$$\begin{bmatrix} h+1-r & 0 & 0 & 0 & \dots & -4 & -4 & \dots \\ 0 & f+1-r & -2 & -2 & \dots & 0 & 0 & \dots \\ 0 & -2 & 4 & 0 & \dots & 0 & 0 & \dots \\ 0 & -2 & 0 & 4 & \dots & 0 & 0 & \dots \\ . & . & . & . & \dots & . & . & \dots \\ -4 & 0 & 0 & 0 & \dots & 16 & 0 & \dots \\ -4 & 0 & 0 & 0 & \dots & 0 & 16 & \dots \\ . & . & . & . & \dots & . & . & \dots \end{bmatrix}.$$

Therefore, the expected values of the diagonal terms c_{ii} and c'_{ii} for the additive and dominance effects from the independent QTL i are $4/N(1-r)$ and $16/N(1-r)$, respectively.

Appendix 2

QTL genotypes falling in marker classes 2 and 3 from backcross 1 and classes 6 and 7 from backcross 2 can not be identified (Moreno-Gonzalez 1993). Thus, the true genotypic values are not known. The mid-value of the two true genotypic values, either a_m or $\frac{1}{2}d_m$, was assigned as the genotypic value of classes 2 and 3; i.e., $\frac{1}{2}a_m + \frac{1}{4}d_m$. Similarly, the assigned genotypic value to classes 6 and 7 is $-\frac{1}{2}a_m + \frac{1}{4}d_m$. Therefore, the square deviation of the assigned values from the true values will contribute to the residual variance by the expression:

$$\Phi = \Sigma r_m(\tfrac{1}{4}a_m^2 + d_m^2/16)$$

where a_m and d_m are the additive and dominance effects, respectively; r_m is the frequency of classes 2, 3, 6 and 7, which is also the recombination frequency between flanking markers of QTL *m*; *m* refers to the QTL in the model. Since the pooled within-backcross genetic variance is

$$\sigma_g^2 = \Sigma(\tfrac{1}{4}a_i^2 + d_i^2/16)$$

where *i* refers to all segregating QTL, then

$$\Phi = r\Sigma_{g''}^2$$

where $\sigma_{g''}^2$ is the pooled within-backcross genetic variance accounted for by QTL already in the model.

References

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

Gill JL (1978) Design and analysis of experiments in the animal and medical sciences, vol 1, Iowa State University Press, Ames, Iowa

Hallauer AR, Miranda Filho JB (1988) Quantitative genetics in maize breeding. Iowa State University Press, Ames, Iowa

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

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124: 743–756

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

Moreno-Gonzalez J (1993) Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques. Theor Appl Genet (in press)

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