Genetic and Nongenetic Bases for the L-Shaped Distribution of Quantitative Trait Loci Effects

Bruno Bost, Dominique de Vienne, Frédéric Hospital, Laurence Moreau and Christine Dillmann

Station de Génétique Végétale, INRA/UPS/INA P-G, F-91190 Gif sur Yvette, France

Manuscript received August 11, 1999
Accepted for publication December 20, 2000

ABSTRACT

The L-shaped distribution of estimated QTL effects (R^2) has long been reported. We recently showed that a metabolic mechanism could account for this phenomenon. But other nonexclusive genetic or nongenetic causes may contribute to generate such a distribution. Using analysis and simulations of an additive genetic model, we show that linkage disequilibrium between QTL, low heritability, and small population size may also be involved, regardless of the gene effect distribution. In addition, a comparison of the additive and metabolic genetic models revealed that estimates of the QTL effects for traits proportional to metabolic flux are far less robust than for additive traits. However, in both models the highest R^2's repeatedly correspond to the same set of QTL.

WITH the use of molecular markers, mapping of loci involved in quantitative variation, *i.e.*, quantitative trait loci (QTL; Geldermann 1975), has progressively become a central method in quantitative genetics. QTL detection is seen as a way to investigate the number of genes that control quantitative traits and the magnitude and the distribution of their effects. In maize, tomato, rice, Drosophila, or mouse, where numerous QTL have been detected for a large number of traits, compilations consistently reveal extremely skewed distributions of QTL effects, with few QTL having large effects, more QTL having moderate effects, and possibly a lot having small effects, resulting in a typical L-shaped distribution. For example, in Drosophila, many loci have small effects on abdominal and sternopleural bristle number, but few loci cause most of the genetic variation (Mackay 1996). Edwards *et al.* (1987) searched for associations between ~20 marker loci and 82 traits in two F_2 populations of maize, each of about 1900 individuals. With a type I error of 5%, they found 2460 significant associations, with a very L-shaped distribution of the fractions of phenotypic variance explained by the detected QTL (R^2). The maximum R^2 value was 16.3%, and 94.5% of the associations exhibited R^2 values <5%, with a minimum at 0.3%. Other examples can be found in the literature (*e.g.*, Sing and Boerwinkle 1987; Shrimpton and Robertson 1988; Paterson *et al.* 1991; Zehr *et al.* 1992; Schön *et al.* 1994; Grandillo and Tanksley 1996; Lee *et al.* 1996; see also Kearsey and Farquhar 1998 for a review in plants).

These L-shaped distributions of QTL effects could partly be due to several statistical artifacts. Most of the results cited above are compilations from several traits and/or several populations/environments and result in a mixture of possibly different distributions. Moreover, QTL effects expressed as fractions of phenotypic variation are proportional to the square of genetic effects, and thus even a normal distribution of true genetic effects would appear skewed toward the smallest values at the R^2 level. Finally, the experimental distributions of QTL effects only concern detected QTL and may be misleading. First, the distribution of estimated R^2 is typically L-shaped when truncated at a given threshold significance value. Second, undetected QTL can inflate the effect of linked detected QTL, as shown very early by McMillan and Robertson (1974). Third, Carbonell *et al.* (1992, 1993) and Beavis (1998) have shown by simulations that the population sizes classically studied (250 individuals for Carbonell *et al.*, 100 for Beavis) lead to both lack of detection of QTL with small effects and possible overestimation of the effects of the detected QTL.

However, a survey of the literature shows that for various traits the same major QTL can be found in different populations or environments, which seems quite unlikely in case of erroneous estimates of the QTL effects. For example, in two different F_2 populations derived from crosses between maize and teosinte, Doebley and Stec (1993) found similar suites of QTL for architectural traits, with the same order of QTL effects for several of them. In three tomato populations derived from the same parents (one F_2 and two $F_{2/3}$) and grown in different environments, Paterson *et al.* (1991) detected QTL for fruit traits that were common to two or even three conditions. Also, apparently the same QTL were

Corresponding author: Christine Dillmann, Station de Génétique Végétale, INRA/UPS/INA P-G, Ferme du Moulon, F-91190 Gif sur Yvette, France. E-mail: dillmann@moulon.inra.fr

detected in different species of Poaceae (Lin *et al.* 1995; Paterson *et al.* 1995), Fabaceae (Fatokun *et al.* 1992; Maughan *et al.* 1996; Timmerman-Vaughan *et al.* 1996), or in distant breeds of animals (Georges and Andersson 1996).

From the biological point of view, there is no reason for a discontinuity between all-or-null (wild-type/mutant) variation and quantitative variation: a continuity between high-effect QTL and low-effect QTL is rather expected. Intermediate cases are found: for example, in pea, a major gene for Ascochyta blight resistance was mapped on chromosome 4 using both a QTL detection approach and a Mendelian analysis after partitioning the distribution of the resistance in the progeny into two classes (Dirlewanger *et al.* 1994). Also, the actual continuous distributions of true genetic effects in a population result from different evolutionary constraints. Theoretical studies (Barton and Turelli 1987; Orr 1999) suggest that L-shaped distributions might be a natural consequence of adaptation toward a fixed optimum.

The issue of this article is to determine whether an apparent L-shaped distribution of estimated QTL effects allows us to make inferences about the true underlying distribution of gene effects, or not. In a recent article, we showed that for any trait proportional to a flux through a linear metabolic pathway at the steady state, the L-shaped distribution is expected as a consequence of the summation property of the control coefficients (Bost *et al.* 1999). In the present article, we study by simulation and analytically other factors likely to influence the distribution of the estimated effects of a set of known QTL controlling an additive trait in a segregating population: the distribution of the additive allelic effects of the underlying genes, linkage disequilibrium, parental gametic phase, environmental effects, and population size. We assumed for simplicity that the genotypes and positions of the loci controlling the quantitative traits are known without error, and restricted our analysis to the estimation of the effects of those genes in an experimental segregating population. The reproducibility of QTL effect estimates was also analyzed. For all the studied factors, we compared the additive model with the metabolic model used by Bost *et al.* (1999).

METHODS

Definitions: Strictly speaking, the "effect" of a locus is classically described by the difference between genetic values of alleles in a population (additive effect). Nevertheless, the "QTL effect" is often referred to in the literature as the fraction of phenotypic variance explained by the QTL (*e.g.*, Edwards *et al.* 1987; Kearsey and Farquhar 1998). To keep with this widely used terminology and be as rigorous as possible, we define here three kinds of effects for a given locus q:

i. The "additive allelic effect" (a_q) is half the difference between the genetic values of "high" and "low" homozygous genotypes at locus q (we assumed no epistasis).
ii. The "true QTL effect" (r_q^2) is the fraction of phenotypic variance explained by the QTL in the model used to decompose the effects,

$$r_q^2 = h_b^2 \frac{\sigma_q^2}{\sigma_G^2},$$

where h_b^2 is the broad sense heritability of the trait, σ_G^2 is the total genetic variance of the trait in the population, and σ_q^2 is the genetic variance contributed by QTL q in the model.
iii. The "estimated QTL effect" (R_q^2) is the statistic that estimates r_q^2 in the experiment

$$R_q^2 = SS_q/\mathrm{SST},$$

where SS_q is the sum of squares corresponding to the QTL q and SST is the total sum of squares.

Model: We consider a cross between two inbred lines and the F_2 population obtained by selfing their F_1 hybrid. In this population, 50 polymorphic QTL determine the value of a quantitative trait, with two alleles, "high" or "low," for each QTL, no dominance, and no epistasis. The genetic value, $\mathcal{G}_{qi}$, of individual i at locus q in the F_2 population is

$$\mathcal{G}_{qi} = m_q + (\theta_{qi} - 1)a_q, \qquad (1)$$

where m_q is the midhomozygote value at locus q, θ_{qi} is the number of high alleles (0, 1, or 2) of individual i at locus q, and a_q is the additive allelic effect at locus q. The genetic value of individual i for the trait is computed under an additive model (ADD) as

$$G_{Ai} = \sum_{q=1}^{50} \mathcal{G}_{qi}, \qquad (2)$$

or under a metabolic model (MET) where the quantitative trait is proportional to a flux through a linear metabolic pathway at the steady state, as

$$G_{Mi} = \frac{K}{\Sigma_{q=1}^{50}[1/\mathcal{G}_{qi}]}, \qquad (3)$$

where K is a constant that characterizes the metabolic pathway (Bost *et al.* 1999). In such a model, the locus q controls the activity of enzyme q, and $\mathcal{G}_{qi}$ depends on the maximal velocity and on the Michaelis constant of enzyme q in individual i (Kacser and Burns 1973).

We compared four distributions of additive allelic effects (a_q) among the loci: (i) constant distribution with all the a_q having the same value; (ii) normal distribution; (iii) exponential distribution with the mode of the distribution corresponding to high a_q values [the probability density function is $f(x) = 1/\sigma \exp[-(x - a_{\max})/\sigma]$, where σ is the standard deviation of the distribution and $a_{\max}$ is the maximal a_q value]; and (iv) uniform distribution, where a_q can take any value between $a_{\min}$ and $a_{\max}$ with the same probability. Whatever the distribution, the m_q are kept constant and identical across loci. To compare the additive and metabolic models, m_q values, as well as the parameters of the a_q distributions, are fitted in the additive model to get approximately the same genetic coefficient of variation as in the metabolic model (Table 1). For each distribution, a fixed sample of 50 a_q values was used in all simulations. The distributions of a_q values in those samples are presented in the first column of Figure 1.

The phenotypic value z_i of each F_2 individual was computed by adding an environmental effect (ε_i) to the genetic value

$$z_i = G_i + \varepsilon_i, \tag{4}$$

where ε_i is randomly drawn from a normal distribution with mean zero and variance σ^2_{E}. σ^2_{E} is computed from the broad sense heritability of the trait (h^2_{b}) and the total genetic variance in the F_2 population (σ^2_{G}):

$$\sigma^2_{\mathrm{E}} = \frac{1 - h^2_{\mathrm{b}}}{h^2_{\mathrm{b}}}\,\sigma^2_{\mathrm{G}}. \tag{5}$$

As defined previously, $R^2_q = SS_q/\mathrm{SST}$, where SS_q is the sum of squares corresponding to the QTL q, and SST is the total sum of squares of the model. With ANOVA, SS_q is computed straightforwardly, and the residuals of the model contain all the variation due to segregation at the other QTL, as well as the random environmental deviation. It is well known that, in this case,

$$R^2_q = \frac{(\kappa - 1)F}{(N - \kappa) + (\kappa - 1)F}, \tag{6}$$

where N is the population size, κ is the number of genotypic classes at QTL q (in an F_2 population, $\kappa = 3$), and F is the test statistics for the effect of QTL q. F follows a noncentral Fisher distribution $\mathcal{F}(\kappa - 1, N - \kappa, \phi_q)$, with

$$\phi_q = (N - 1)\frac{r^2_q}{1 - r^2_q}$$

being the noncentrality parameter (Sheffé 1959). Note that the QTL effect (R^2) estimation is biased and overestimates the true QTL effect (r^2) in small samples (Charcosset and Gallais 1996). R^2 was preferred here rather than the unbiased "adjusted R square" because it has a lower sampling variance (Cramer 1987).

TABLE 1

Genetic parameters used in the simulations

Distribution law for the additive allelic effects	Genetic parameters[a]	Parameters of the law	Model	
			ADD[b]	MET[c]
Constant	m_q ($\forall q$)		0.1017	20.00
	a_q ($\forall q$)		0.0512	10.00
Normal	m_q ($\forall q$)		0.1000	20.00
	a_q	μ	0.0512	10.00
		σ	0.0126	2.50
Exponential[d]	m_q ($\forall q$)		0.1000	20.00
	a_q	$a_{\max}$	0.0918	16.20
		σ	0.0126	2.50
Normal	m_q ($\forall q$)		0.1000	20.00
	a_q	$a_{\min}$	0.0136	2.50
		$a_{\max}$	0.0950	17.50

[a] m_q, the midhomozygote value at locus q; a_q, the additive allelic effect at locus q.

[b] Additive model.

[c] Metabolic model.

[d] The probability density function for the exponential distribution is

$$f(a_q = x) = \frac{1}{\sigma}\exp\left[-\frac{x - a_{\max}}{\sigma}\right].$$

When several QTL are taken simultaneously into account via multiple regression, a relationship similar to the one in Equation 6 is found between the global R square of the model and the corresponding F-statistics. However, the partial sum of squares for QTL q (SS^*_q) takes into account all the other QTL of the model, and the SST is not equal to the sum of all SS^*_q plus the residual sum of squares (SSR*; see appendix a for details). Hence, there is no simple relationship between F and R^2_q, and the theoretical distribution of R^2_q values is unknown.

Genetic maps and parental genotypes: Linkage disequilibrium between QTL depends on both the genetic map and the genotype of the parental inbred lines. We considered four different genetic structures for the pairs of parental inbred lines. The pair *RandomU* had independent QTL and random gametic phase: the genotype of one parent at each locus was drawn at random, so that the parental gametic phase (succession of high and low alleles along the genome) was random; the other parent had the complementary succession of high and low alleles, since all the loci are polymorphic. The pair *RandomL* had linked QTL and random parental gametic phase. The pair *CouplingL* had linked QTL that were in coupling, with one parent having the low alleles at all loci and the other parent having the high alleles at all loci. The pair *RepulsionL* had linked QTL that were in repulsion, with an alternation of high and low alleles along the chromosomes in each parent. The same genetic map (set of QTL locations on chromosomes) was used for all the simulations involving linked QTL, where QTL locations were randomly spread over 10 chromosomes of 200 cM each.

Simulations: For each situation, we performed Monte Carlo simulations to obtain 100 replicates of the F_2 population. Each replicate consisted of the following sequence:

TABLE 2

Statistics describing the simulated distributions of the R^2

Linkage	Parents	h_b^2	Allelic effects	Population size	PCT (%)	Skew	$\bar{S}$ (%)	W	N2	HPF (%)
				Independent QTL, $h_b^2 = 1.0$						
No	*RandomU*	1.0	Constant	200	100	0.51	64.3	0.438	23	4
No	*RandomU*	1.0	Constant	1000	74	0.10	96.3	*ns*	49	6
No	*RandomU*	1.0	Constant	5000	57	0.06	99.6	*ns*	48	5
No	*RandomU*	1.0	Normal	200	76	0.79	76.5	0.964	5	33
No	*RandomU*	1.0	Normal	1000	59	0.64	95.2	0.995	3	100
No	*RandomU*	1.0	Exponential	1000	46	−0.69	96.3	0.985	10	42
No	*RandomU*	1.0	Uniform	1000	58	0.53	95.7	0.999	4	80
				Linked QTL, $h_b^2 = 1.0$						
No	*RandomL*	1.0	Constant	200	96	0.70	44.0	0.954	9	35
No	*RandomL*	1.0	Constant	500	88	0.61	51.5	0.978	5	42
No	*RandomL*	1.0	Constant	1000	85	0.59	54.4	0.987	4	43
No	*RandomL*	1.0	Constant	5000	83	0.57	56.4	0.995	4	57
No	*RandomL*	1.0	Normal	1000	88	1.15	50.3	0.993	2	96
No	*RandomL*	1.0	Exponential	1000	86	0.67	52.7	0.0994	3	99
No	*RandomL*	1.0	Uniform	1000	88	0.87	43.5	0.994	2	98
No	*CouplingL*	1.0	Constant	1000	100	0.59	14.0	0.987	4	45
No	*RepulsionL*	1.0	Constant	1000	53	0.58	125.5	0.987	4	45
				Independent or linked QTL, $h_b^2 < 1.0$						
No	*RandomU*	0.2	Constant	200	91	0.73	36.3	*ns*	49	2
No	*RandomU*	0.2	Constant	1000	100	0.47	23.3	*ns*	48	5
No	*RandomU*	0.5	Constant	200	86	1.47	50.2	*ns*	49	2
No	*RandomU*	0.5	Constant	1000	97	0.69	50.0	*ns*	48	6
No	*RandomU*	0.2	Normal	1000	99	1.39	23.3	0.195	42	3
No	*RandomU*	0.5	Normal	1000	92	0.95	50.8	0.506	20	7
Yes	*RandomL*	0.2	Constant	1000	100	0.30	14.9	0.203	38	9
Yes	*RandomL*	0.5	Constant	1000	99	1.20	29.8	0.540	24	14

The R^2 distributions were computed for a 50-locus additive trait, in 100 F_2 populations, and with four distributions of allelic effects (a_q): (i) Constant (same a_q for all the QTL), (ii) Normal, (iii) Exponential, and (iv) Uniform. The QTL were independent or linked, with various parental gametic phases. h_b^2 is the broad sense heritability of the trait. See METHODS for the details on parameters PCT, Skew, $\bar{S}$, W, N2, and HPF.

N F_2 individuals were drawn at random by selfing the F_1 hybrid ($N = 200, 500, 1000, 5000$), assuming recombinations with no interference (Hospital and Chevalet 1993).

The genetic values G_{Ai} or G_{Mi} of all the F_2 individuals were computed. The total genetic variance (σ_G^2) was then computed, allowing us to compute the environmental variance (σ_E^2) from the given broad sense heritability (h_b^2), following Equation 5. Then the random environmental deviations were added to get phenotypic values of the F_2 individuals.

Finally, multiple regression of the phenotypic values was directly performed on the genotype at each of the 50 QTL. The fraction of the phenotypic variation explained by the qth QTL in the kth sample of a given F_2 population was estimated as $R_{qk}^2 = SS_{qk}/\mathrm{SST}_k$. Sums of squares were computed with SAS GLM procedures (SAS Institute 1988).

Distribution of estimated QTL effects (R_q^2): The $50 \times 100 = 5000$ R_{qk}^2 values obtained for each genetic model were used to characterize the distribution of estimated QTL effects. The shape of the distributions was first characterized by their skewness, Skew (Sokal and Rohlf 1995). With 50 underlying genes having the same additive allelic effect in an infinite population, and no environmental variation ($h_b^2 = 1$), each QTL should explain $1/50 = 2\%$ of the total variation. Thus we also computed PCT, which is the cumulative frequency of the distribution corresponding to $R^2 = 2\%$.

As each QTL is known in our model, we should explain 100% of the phenotypic variation by adding the effects of each QTL. Thus we computed the proportion of the phenotypic variation explained by the model, averaged over the 100 replicates ($\bar{S}$):

$$\bar{S} = \frac{1}{100}\sum_{k=1}^{100}\left[\sum_{q=1}^{50} R_{qk}^2\right].$$

Finally, to measure the repeatability of the ranking of the estimated QTL effects among the replicates, we ranked the 50 QTL in each replicate according to their R^2 values and computed three parameters: (i) Kendall's

TABLE 3

Comparison between additive and metabolic models

Simulation parameters			Skew		W		N2		HPF (%)	
h_b^2	Population size	Allelic effects	ADD[a]	MET[b]	ADD	MET	ADD	MET	ADD	MET
		Independent QTL (*RandomU*)								
$h_b^2 = 1.0$	$N = 1000$	Constant	0.10	0.29	NS	NS	49	49	6	6
$h_b^2 = 1.0$	$N = 1000$	Normal	0.64	3.09	0.995	0.905	3	4	100	100
		Linked QTL (*RandomL*)								
$h_b^2 = 1.0$	$N = 5000$	Constant	0.57	0.63	0.995	0.944	4	11	57	31
$h_b^2 = 1.0$	$N = 1000$	Constant	0.59	0.84	0.987	0.773	4	18	43	27
$h_b^2 = 1.0$	$N = 200$	Constant	0.70	1.38	0.954	0.334	9	28	35	18
$h_b^2 = 0.5$	$N = 1000$	Constant	1.20	1.33	0.540	0.409	24	27	14	10
$h_b^2 = 0.2$	$N = 1000$	Constant	1.88	1.85	0.203	0.145	38	40	9	11
$h_b^2 = 1.0$	$N = 1000$	Normal	1.15	2.10	0.993	0.817	2	6	96	52

The R^2 distributions were computed for a 50-locus additive or metabolic trait, in 100 F_2 populations, and with two distributions of additive allelic effects (a_q): Constant (same a_q for all the QTL) and Normal. The QTL were independent or linked, with random parental gametic phase. h_b^2 is the broad sense heritability of the trait. See METHODS for the details on parameters Skew, W, N2, and HPF. NS, not significant.

[a] Additive model.

[b] Metabolic model.

concordance coefficient, W (Kendall 1955), between replicates; (ii) the number, N2, of different QTL that appeared in at least one replicate among the two QTL displaying the highest R^2 values; and (iii) HPF is defined so that the QTL, which was the most often first ranked, was first ranked in HPF% of the replicates. A W value close to 1 indicates that the ranking of the QTL is similar over the replicates. In this case, N2 would be equal to 2, and HPF would be equal to 100%. The maximum value of N2 is 50, the total number of QTL. These three parameters are different ways to measure the repeatability of the ranking of the QTL: W takes into account all the 50 QTL, N2 corresponds only to the two highest QTL in each replicate, whereas HPF is concerned only with the highest QTL. The values obtained for those parameters under the different situations are given in Tables 2 and 3.

RESULTS

Distribution of estimated QTL effects: We used simulations to study the factors likely to influence the distribution of the estimated effects (R^2) of a set of 50 known QTL controlling an additive trait in an F_2 population: distribution of the additive allelic effects (a_q) of the underlying genes, linkage disequilibrium, parental gametic phase (coupling or repulsion), environmental effects, and population size. In addition to the simulations, we performed analytical calculations, taking into account simpler situations, to explain the mechanisms that are involved for the differents factors. The details of these analytical calculations are given in APPENDICES A and B.

Independent QTL, without environmental variation: In that case, the only source of variation for a given QTL is the sampling of individuals in the population, which leads to some deviations of R^2 around the true QTL effect r^2. At each QTL, the observed proportions of the three different genotypes randomly deviate from the theoretical (1:2:1) proportions. It is easy to show, using the notations introduced in APPENDIX B, that this sampling of individuals leads to an underestimation of the variance contributed by the QTL,

$$\hat{\sigma}_q^2 = \sigma_q^2 - 2(\hat{f}_0 - \hat{f}_2)^2 a_q^2 ,$$

where $\hat{f}_0$ and $\hat{f}_2$ are the observed frequencies of the homozygote genotypes, and σ_q^2 is the genetic variance contributed by QTL q in the infinite F_2 population. As shown in Table 2, a consequence is that the average sum of R^2 ($\bar{S}$) is always <100%, even with a population of 1000 individuals (between 95.2 and 96.3%, Table 2). $\bar{S}$ increases when population size increases and tends toward 100% for a population of 5000 individuals.

When underlying genes have identical additive allelic effects, the true effect (r^2) is obviously the same for all the QTL. As shown in Figure 1 (Constant, *RandomU*), the random deviations of R^2 around r^2 are moderate with $N = 1000$ individuals. When underlying genes have nonidentical effects, the distribution of R^2 values roughly corresponds to the distribution of true QTL effects (Figure 1, *RandomU*).

Linked QTL, without environmental variation: When the QTL are randomly located over the genome, linkage between QTL occurs, which consistently results in L-shaped distributions of R^2 values, regardless of the distribution of additive allelic effects (Figure 1, *RandomL*). The data in Table 2 confirm this tendency: the values of skewness (Skew) are high and positive, and

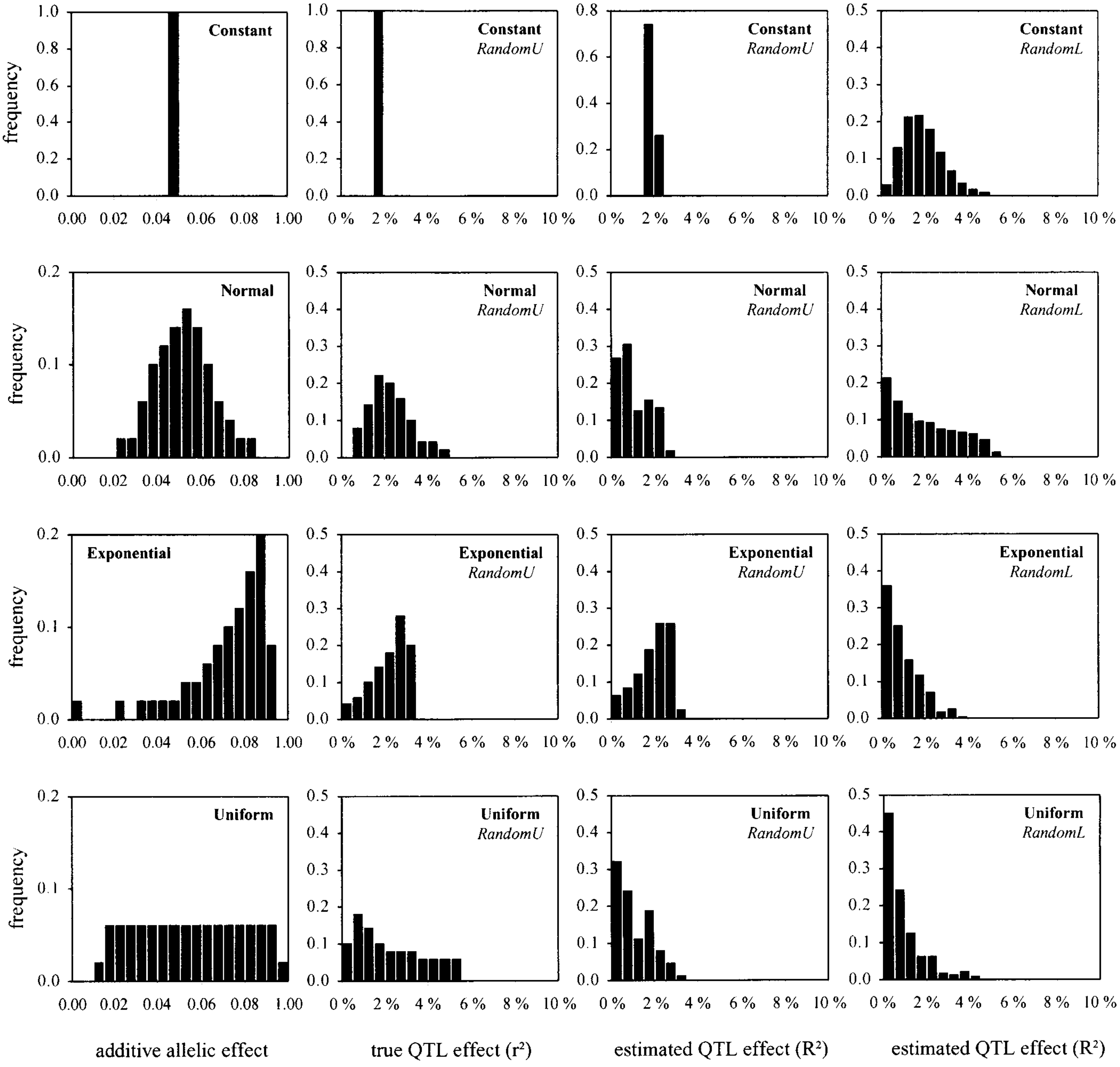

FIGURE 1.—Distribution of additive allelic effects (a_q, first column), true QTL effects, (r_q^2, second column), and corresponding distributions of estimated QTL effects (R_q^2, third and fourth columns) obtained in the simulations with the additive model and with two different parental inbred lines: *RandomU*, independent QTL; *RandomL*, linked QTL with random gametic phase. The broad sense heritability of the trait was $h_G^2 = 1$, and the size of the F_2 population was $N = 1000$. The parameters used for the distributions of the allelic additive effects are given in Table 1. The corresponding r_q^2 values were obtained for independent QTL in an infinite F_2 population.

PCT values are showing that most (between 53 and 100%) of the R^2 values are <2%.

To explain these observations, we calculated the values of true QTL effects (r^2) in the simple case of three linked QTL (Q_1, Q_2, and Q_3; see APPENDIX B for details). Whether the detection method is one-way ANOVA or multiple regression as used in our simulations, the r^2 values appear to depend on both additive allelic effects (a_{Q_1}, a_{Q_2}, and a_{Q_3}) and linkage disequilibria (D_{12}, D_{23}, and D_{13}) between QTL Q_1, Q_2, and Q_3 in the F_2 population. The total genetic variance in the F_2 population (σ_G^2) is

$$\sigma_G^2 = \frac{1}{2}(a_{Q_1}^2 + a_{Q_2}^2 + a_{Q_3}^2) + 4D_{12}\, a_{Q_1}\, a_{Q_2} + 4D_{23}\, a_{Q_2}\, a_{Q_3} + 4D_{13}\, a_{Q_1}\, a_{Q_3} \qquad (7)$$

(APPENDIX B). Thus, for given values of the additive allelic effects, the total genetic variance increases with coupling ($D > 0$) and decreases with repulsion ($D < 0$).

With ANOVA, we have

$$r^2_{Q_1} = \frac{\frac{1}{2}(a_{Q_1} + 4D_{12}\,a_{Q_2} + 4D_{13}\,a_{Q_3})^2}{\sigma^2_G} \quad (8)$$

$$r^2_{Q_2} = \frac{\frac{1}{2}(a_{Q_2} + 4D_{12}\,a_{Q_1} + 4D_{23}\,a_{Q_3})^2}{\sigma^2_G} \quad (9)$$

$$r^2_{Q_3} = \frac{\frac{1}{2}(a_{Q_3} + 4D_{23}\,a_{Q_2} + 4D_{13}\,a_{Q_1})^2}{\sigma^2_G} \quad (10)$$

(appendix b). Thus, the genetic variances contributed by linked QTL (numerators of Equations 8, 9, and 10 increase with coupling and decrease with repulsion.

With multiple regression, we have

$$r^2_{Q_1} = \frac{\frac{1}{2}a^2_{Q_1}}{\sigma^2_G}(1 - 16D^2_{12}) \quad (11)$$

$$r^2_{Q_2} = \frac{\frac{1}{2}a^2_{Q_2}}{\sigma^2_G}\left[\frac{(1 - 16D^2_{12})(1 - 16D^2_{23})}{1 - 256D^2_{12}D^2_{23}}\right] \quad (12)$$

$$r^2_{Q_3} = \frac{\frac{1}{2}a^2_{Q_3}}{\sigma^2_G}(1 - 16D^2_{23}) \quad (13)$$

(appendix b). Thus, in the multiple regression model, whatever the gametic phase (coupling or repulsion), linkage disequilibrium decreases the genetic variance contributed by linked QTL (numerators of Equations 11, 12, and 13). In this case, if all the additive allelic effects are identical, the QTL with the highest r^2 are the independent ones ($D = 0$). However, the sum of individual r^2 still depends on the gametic phase. From Equations 7 and 11–13, it is expected to be <100% in case of coupling ($D > 0$) and >100% in case of repulsion ($D < 0$). This is consistent with our simulations when we compare the sum, $\overline{S}$, of the estimated QTL effects (R^2) between *CouplingL* and *RepulsionL* parental inbred lines (Table 2). With equal additive allelic effects (constant), the *CouplingL* gives $\overline{S} = 14.0\%$, while the *RepulsionL* gives $\overline{S} = 125.5\%$. The sum is intermediate ($\overline{S} \simeq 50.0\%$) with *RandomL* parental inbred lines.

In general, the genetic variance of a QTL q should also depend on linkage disequilibria between QTL q and all the QTL linked to q, or linkage disequilibria of higher order. However, if the genetic variance contributed by the QTL q is computed by multiple regression, we showed that it depends only on the first order linkage disequilibria between QTL q and its nearest neighbors, the QTL $q - 1$ and $q + 1$ (Equations 11 and 13), in the case of an F_2 population. Thus, the flanking QTL tend to absorb the effects of all nearby QTL. Stam (1991) showed a similar property of multiple regression when applied to QTL detection in backcross populations: the two flanking markers of a QTL tend to absorb the effects of all nearby QTL.

In our simulations, we investigated the relationship between R^2 and the squared linkage disequilibrium value between each QTL and its two nearest neighbors, in the F_2 population resulting from the cross between *RandomL* parental inbred lines, and with identical additive allelic effects for all the 50 QTL. A nonlinear multiple regression of R^2 values on squared linkage disequilibria was performed on these data to check Equation 12. The regression showed that the simulation results perfectly fit the analytical results: the determination coefficient of the regression is 99.9%, and the parameter estimates are very close to the theoretical coefficients in Equation 12. This result is illustrated in Figure 2. This observation confirms that the r^2 value of a given QTL depends only on pairwise linkage disequilibria between the QTL and its nearest neighbors. Hence, unequal map distances between QTL, which result in an L-shaped distribution of squared linkage disequilibria between QTL, should contribute to the L-shaped distribution of the estimated QTL effects. On the other hand, as QTL effects are confounded with the genetic distances between the QTL and its two neighbors, equally spaced QTL should result in two different R^2 values: one for QTL flanked by two other QTL and one for QTL at the beginning or the end of the chromosomes. Simulations with equally distributed distances between QTL showed indeed that all the QTL have the same R^2 values, once the QTL located at the ends of the chromosomes were excluded (not shown). In the general case, the true QTL effects also depend on the additive allelic effects of each QTL. Relying on Equations 11–13, we expect QTL with the highest r^2 in a given population to be either independent QTL or QTL with high additive allelic effect. This point was confirmed by checking the R^2 values of individual QTL in the simulations (not shown).

Environmental variation and reduced population size: When there is environmental variation on the trait ($h^2_b < 1$), or small sample size, the distribution of R^2 is again L-shaped, whatever the distribution of additive allelic effects, and whether there is linkage or not: all Skew values are significant and positive, and PCT values are between 86 and 100% (Table 2).

To explain this phenomenon, we considered the simplest case of one-QTL analysis of variance (appendix a). Using Equation 6, we performed numerical computations of the distribution of R^2 from the distribution of F. We consider Q independent QTL with the same additive allelic effect on an additive trait. The total genetic variance is $\sigma^2_G = Q\sigma^2_q$, and the percentage of phenotypic variance explained by each QTL is $r^2_q = h^2_b/Q$. Thus the noncentrality parameter for the distribution of F is

$$\phi = (N - 1)\,\frac{h^2_b}{Q - h^2_b}\,.$$

For each R^2 value, the probability $P(X \leq R^2) = P(Y \leq$

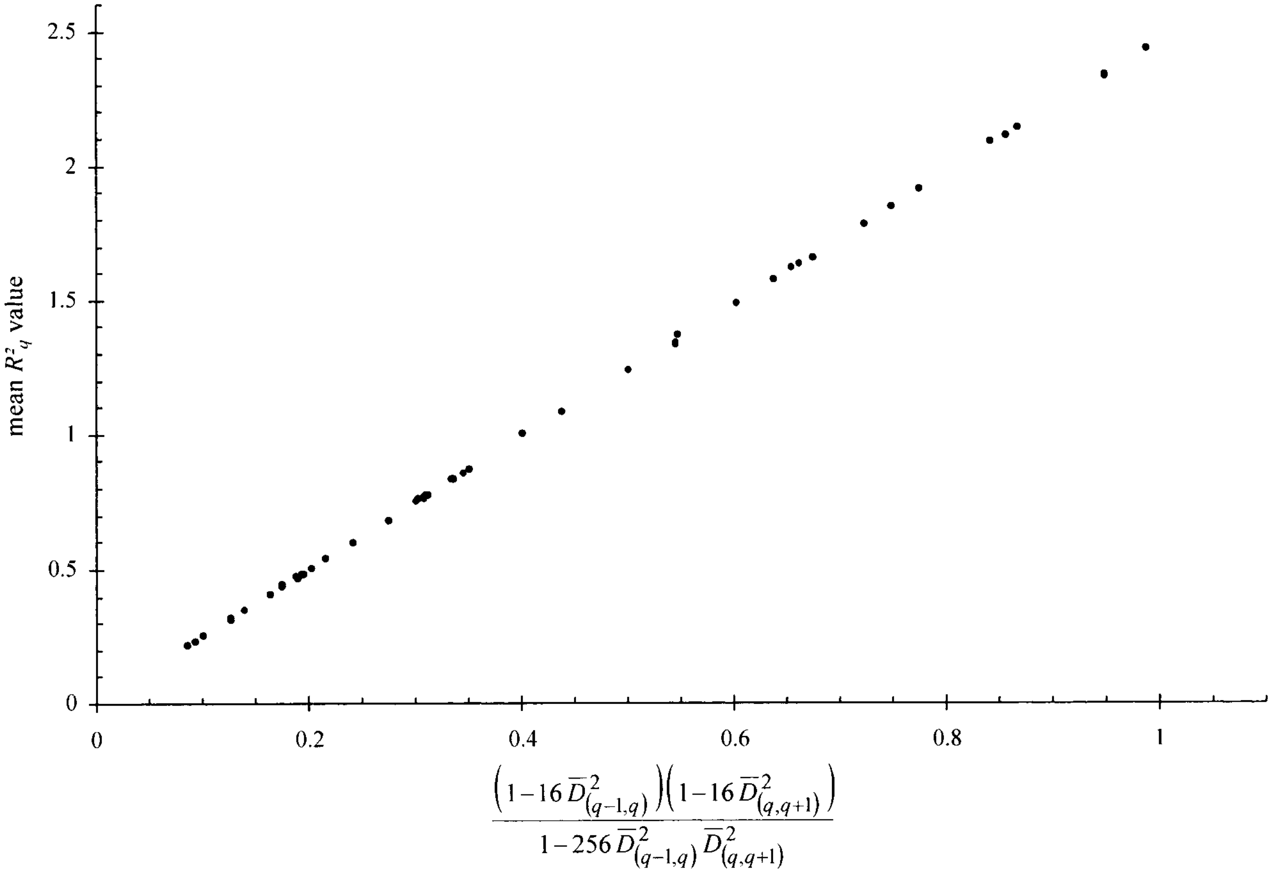

FIGURE 2.—Relationship between estimated QTL effects (R^2) and pairwise linkage disequilibria. For each QTL q, R^2_q values were averaged over the 100 simulation replicates. Linkage disequilibria $D_{q-1,q}$ and $D_{q,q+1}$ between QTL q and its two flanking neighbors ($D = 0$ for QTL located at the end or at the beginning of a chromosome) were computed from the genetic distances between the QTL. We considered the case where the 50 QTL were randomly linked (*RandomL*) and had the same additive allelic effect (Constant), with $h^2_b = 1.0$ and $N = 1000$. $(1 - 16D^2_{q-1,q})(1 - 16D^2_{q,q+1})/(1 - 256D^2_{q-1,q}\,D^2_{q,q+1})$ is the theoretical linkage disequilibrium coefficient of r^2_q in the three-QTL model (Equation 12).

F) was computed to get the theoretical distribution of ANOVA R^2 values (open bars in Figure 3) for various heritabilities ($h^2_b = 0.2, 0.5, 1.0$) and various population sizes ($N = 200$ and 1000). The resulting distributions were compared to those obtained by simulation with multiple regression (solid bars in Figure 3), taking the same values for the heritability (h^2_b) and population size (N). With ANOVA, the sampling distribution of R^2 depends on both parameters, N and h^2_b (CRAMER 1987). As h^2_b decreases, ϕ decreases as well as r^2 and the average $\overline{R}^2$, but the sampling variance of R^2 increases. On the contrary, the population size does not influence r^2, but as N decreases, the sampling mean and sampling variance of R^2 increase.

The higher the sampling variance of R^2, the more pronounced the L-shaped distribution of QTL R^2 values. An intuitive explanation for such a distribution shape is that, with random errors, the most likely event is that the intraclass variability hides the difference between genotypic classes at one QTL, leading to R^2 values low or close to zero. However, it may occur, by chance, that random samples of individuals (genotypes) or environments reinforce the differences between genotypic classes and lead, for some QTL, to R^2 values $>r^2$. The higher the r^2, the lower the chance of such event. With ANOVA, the L-shaped distribution occurs in small populations ($N = 200$), even with $h^2_b = 1$, because of the segregation at the other QTL. For this reason, the distribution is very sensitive to the population size. With multiple regression all known QTL are taken into account in the model and the L-shaped distribution of R^2 mainly occurs because of environmental noise. Thus, the L-shaped distribution is not observed with multiple regression in small populations when $h^2_b = 1$. As the heritability of the trait decreases, the differences between ANOVA and multiple regression decrease (Figure 3).

Comparing ranking order of QTL effects in different samples of a population: Given that the distribution of estimated QTL effects (R^2_q) may be quite different from the distribution of the corresponding additive allelic effects (a_q), the question of the reproducibility of the ranking order of the R^2_q in independent experiments performed from a given population arises.

Independent QTL without environmental variation: The

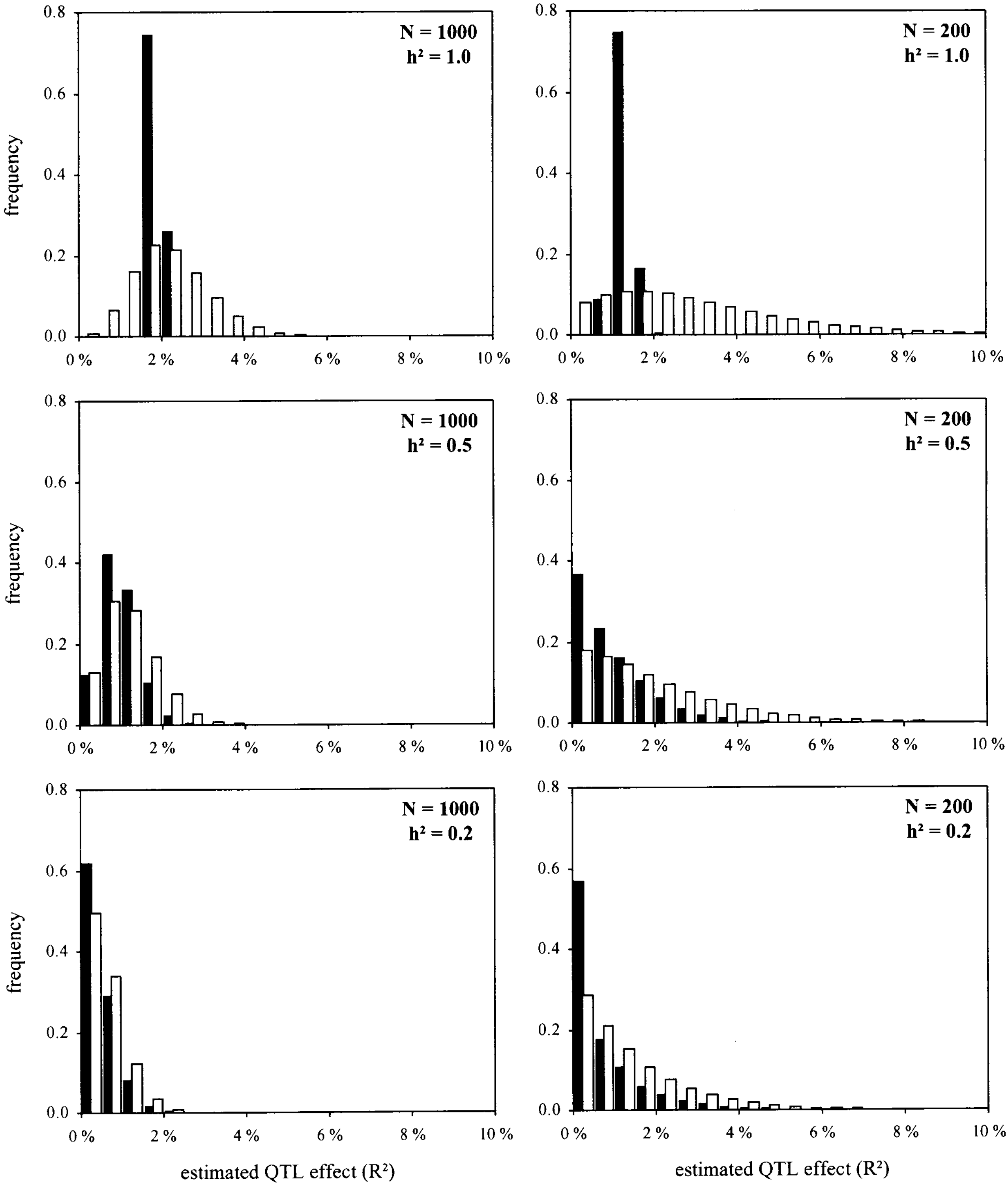

FIGURE 3.—Effect of environmental variation and population size on estimated QTL effects (R^2) distributions. We compared theoretical R^2 distributions obtained for one-QTL analysis of variance (open bars) and R^2 distributions obtained in simulations with multiple regression (solid bars). The ANOVA distributions were obtained using the relationship between R^2 and the test statistics F of the QTL effect (Equation 6 in METHODS). We considered the case where the 50 QTL were independent (*RandomU*) and had the same additive allelic effect (Constant), with different environmental variances (h_0^2 = 1.0, 0.5, 0.2) and population sizes (N = 1000 and 200).

parental inbred lines are *RandomU* and the heritability is $h_b^2 = 1$. When genes have identical allelic effects, there is of course no correlation between replicates (Table 2). On the contrary, when genes have unequal effects, the distributions of r^2 and R^2 reflect the distribution of additive allelic effects. Hence, the ranking of the QTL is well conserved between replicates: W values between 0.985 and 0.999 for populations of 1000 individuals, N2 is between 3 and 10, and HPF is between 42 and 100% (Table 2).

Linked QTL, without environmental variation: The parental inbred lines are *RandomL*, *CouplingL*, or *RepulsionL* and the heritability is $h_b^2 = 1$. In this case, the distributions of r^2 and R^2 reflect both the additive allelic effects and linkage disequilibria between QTL. Hence, the ranking of the QTL is again well conserved between replicates because of unequal linkage disequilibrium distribution (Table 2). However, a gene with a large allelic effect may repeatedly be found as a QTL with a small R^2 if this gene is close to other genes (Equation 9).

Environmental variation and reduced population size: QTL are independent (*RandomU*) or linked (*RandomL*). As expected, whatever the distribution of r^2, the correlation between the rankings of the R^2's decreases when heritability decreases. This correlation is close to the heritability of the trait, since heritability describes the quality of the prediction of genetic values by phenotypic values. For a given heritability, the rank correlation decreases with the population size.

Comparison between additive and metabolic models: In a previous article, we studied the distribution of r^2 for a trait related to a metabolic flux (Bost *et al.* 1999). An L-shaped distribution was consistently found with unequal additive allelic effects, due to unequal flux control coefficients between the parents. However, the distribution of the estimated R^2 under MET is quite similar to the one obtained with ADD when other factors such as linkage disequilibrium, low heritability, or low population size influence the estimation. A comparison between the two genetic models is given in Table 3.

When all genes have the same additive allelic effect, the distribution of R^2 is slightly more skewed with MET than with ADD. The differences are greater when the population size is smaller. Even with unequal additive allelic effects, the differences between MET and ADD are reduced when QTL are linked, as compared to independent QTL.

Actually, the main feature of MET is that correlation between the rankings of the QTL in different replicates is lower than for ADD, in particular for small population sizes (Table 3): at least 5000 individuals are needed with MET to obtain rank correlations similar to those observed for ADD with only 200 individuals. In fact, the variances of each R_q^2 over replicates are always ∼10 times larger with MET than with ADD (not shown). Such low correlations between replicates can be explained by dominance and epistatic effects, which are inherent to MET (Kacser and Burns 1981). However, with MET, the highest R^2 repeatedly corresponded to the same set of QTL, as shown by N2 and HPF values (Table 3). Even when HPF is ∼50%, we checked that only two or three QTL are ranked first in the 100 replicates. The rank correlation between replicates also decreases with the heritability of the trait, though the correlation is always slightly lower than the heritability. But, with MET, the robustness of estimated QTL effect is approximately equally sensitive to heritability and population size.

DISCUSSION

We analyzed different factors likely to influence the estimates of the effects of a finite set of QTL, whose "real" individual effects and positions are known. Even though in experimental situations the number of *detected* QTL is not very high for a given trait (usually <10 with common sizes of population), the numbers of QTL really contributing to the variation of complex traits are expected to be much higher. Accordingly, a set of 50 QTL was chosen for the simulations. The question of mapping these QTL by the use of molecular markers was not considered, because it is analyzed in length in other articles, and because we focused on the influence of particular genetic and nongenetic factors on the distribution of estimated QTL effects.

QTL effects are generally estimated by the statistic R^2, which measures the fraction of phenotypic variation explained by the variation at the QTL. Our simulations have shown that the distribution of the estimated QTL effects (R^2) reflects the distribution of additive allelic effects only if several conditions are combined: no linkage disequilibrium, linear relationship between gene effects and trait values (additive model), no environmental variance, and large population size. Otherwise the R^2 distributions are clearly L-shaped, even with an additive model and J-shaped distributions of additive allelic effects (exponential). We showed here that the distribution of R^2 depends on both genetic and nongenetic factors. Genetic factors such as the distribution of additive allelic effects, the model (additive or metabolic), the linkage disequilibrium between QTL, or the parental gametic phase determine the distribution of r^2, the "true" QTL effects. Then, for each QTL, the R^2 value is a random variable, which depends of course on r^2, but also on the amount of residual noise that confuses the estimation of r^2.

Among the factors likely to increase the residual noise, the heritability of the trait and the population size do not play the same role. The heritability of the trait buffers the relationship between phenotypic and genetic values. It is of course always possible to enhance the heritability of a trait by using refined experimental designs involving progeny tests or cloning. The population size determines the amount of sampling variation for the genotypic composition of the experimental population. In a given genotypic class at a given QTL, the latter mainly affects the segregation at the other QTL. The resulting overestimation of detected QTL effects and lack of repeatability with small population sizes have already been documented by Carbonnel *et al.* (1992, 1993) and Beavis (1998). When all QTL are known, as was assumed in our simulations, these sampling effects due to environmental noise and reduced population size could be partly minimized via multiple regression. The only case where population size really becomes prejudicial is when the genetic effect of the QTL depends on the genetic background, *i.e.*, with epistasis, as in the metabolic model. However, it is worth noting that the highest R^2 values repeatedly correspond to the same small set of QTL, even if the ranks could be modified within this group.

Another cause for the L-shaped distribution of the R^2 values is the L-shaped distribution of r^2 values themselves. Unequal linkage disequilibrium between QTL appears to be a cause for the L-shaped distribution of the r^2 values. We showed analytically that when using one-QTL analysis of variance for estimating the QTL effects, coupling in the parents increases the r^2 values, while repulsion decreases them. When using multiple regression, analytical developments as well as simulations showed that linkage disequilibrium decreases the r^2 values, as a function of the squared linkage disequilibria, whatever the gametic phase and the additive effects of the nearby QTL. Beyond the estimation methods, it is clear that the R^2 values reflect not only the additive allelic effects but also the relative position of the QTL on the genetic map. It is likely that the linkage disequilibrium bias occurs in actual situations and that it is unequally distributed. As unequal map distances between QTL would result in an L-shaped distribution of squared linkage disequilibria between QTL, linkage disequilibrium would therefore contribute to the L-shaped distribution of R^2.

For some applications of QTL methodology, for example, in the context of the candidate gene approach, it is important to get accurate estimates of the effects/positions of the QTL. Our results emphasize that the multiple regression or composite interval mapping methods, which take into account the other QTL as cofactors (Stam 1991; Jansen 1993; Rodolphe and Lefort 1993; Jansen and Stam 1994; Zeng 1994), should be preferred to classical detection methods (ANOVA and simple interval mapping), because they minimize the effect of linkage disequilibrium.

However, these methods have also some drawbacks. First, only detectable QTL can be used as cofactors in the multiple regression, and the residual variation is still expected to include the genetic variation at unknown QTL. Second, linkage disequilibria between markers and QTL, as well as the apparent effect of several QTL located in the same interval between two markers, will influence the r^2 values. Thus, even with these methods, we can predict neither the distribution of the true effects of the QTL (r^2) nor the distribution of additive allelic effects from the estimated effects (R^2). But here the question of which parameter is the most relevant to describe the effect of a gene arises—its true effect or its additive allelic effect?

Besides the statistical tools, different methods may be used to confirm the presence of a QTL in a given chromosomal region. First, QTL detection may be done in different genetic backgrounds, *i.e.*, with different gametic phases. Second, the fine mapping methods, using near isogenic lines or introgression lines (Paterson *et al.* 1990; Eshed and Zamir 1995), highly recombinant inbred lines (Liu *et al.* 1996), or populations possibly panmictic and genetically isolated (*e.g.*, Weeks and Lathrop 1995, for a review) will make it possible not only to map QTL within intervals of 1–2 cM or less but also to estimate their effects with minimal influence of the nearby QTL thanks to minimal linkage disequilibrium. Such approaches are *sine qua non* to have access to actual distributions of QTL controlling quantitative traits and to determine the generality of the L-shaped distribution expected for the traits proportional to metabolic flux, *i.e.*, putatively numerous quantitative traits.

We are very grateful to A. Leonardi for helpful discussions and reading the manuscript, and also to P. Keightley and anonymous referees for helpful comments on the manuscript. B. Bost was supported by a Ph.D. grant from the French Ministry of National Education, Research, and Technology (MENRT).

LITERATURE CITED

Barton, N. H., and M. Turelli, 1987 Adaptative landscapes, genetic distances and evolution of quantitative characters. Genet. Res. **49:** 157–173.

Beavis, W. D., 1998 QTL analyses: power, precision and accuracy, pp. 145–162 in *Molecular Dissection of Complex Traits*, edited by A. H. Paterson. CRC Press, Boca Raton/New York.

Bost, B., C. Dillmann and D. de Vienne, 1999 Fluxes and metabolic pools as model traits for quantitative genetics. I. L-shaped distribution of gene effects. Genetics **153:** 2001–2012.

Carbonell, E. A., T. M. Gerig, E. Balansard and M. J. Asins, 1992 Interval mapping in the analysis of nonadditive quantitative trait loci. Biometrics **48:** 305–315.

Carbonell, E. A., M. J. Asins, M. Balsega, E. Balansard and T. M. Gerig, 1993 Power studies in the estimation of genetic parameters and the localization of quantitative trait loci for backcross and doubled haploid populations. Theor. Appl. Genet. **86:** 411–416.

Charcosset, A., and A. Gallais, 1996 Estimation of the contribution of quantitative trait loci (QTL) to the variance of quantitative traits by means of genetic markers. Theor. Appl. Genet. **93:** 1193–1201.

Cramer, J. S., 1987 Mean and variance of R^2 in small and moderate samples. J. Econometrics **35:** 253–266.

Dirlewanger, E., P. G. Isaac, S. Ranade, M. Belajouza, R. Cousin *et al.*, 1994 Restriction fragment length polymorphism analysis of loci associated with disease resistance gene and developmental traits in *Pisum sativum* L. Theor. Appl. Genet. **88:** 17–27.

Doebley, J., and A. Stec, 1993 Inheritance on the morphological differences between maize and teosinte: comparison of results for two F_2 populations. Genetics **134:** 559–570.

Edwards, M. D., C. W. Stuber and J. F. Wendel, 1987 Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116:** 113–125.

Eshed, Y., and D. Zamir, 1995 An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield associated QTL. Genetics **141:** 1147–1162.

Fatokun, C. A., D. I. Menancio-Hautea, D. Danesh and N. D. Young, 1992 Evidence for orthologous seed weight genes in cowpea and mung bean based on RFLP mapping. Genetics **132:** 841–846.

Geldermann, H., 1975 Investigations on inheritance of quantitative characters in animal by gene markers. I. Methods. Theor. Appl. Genet. **46:** 319–330.

Georges, M., and L. Andersson, 1996 Livestock genomics comes of age. Genome Res. **6:** 907–921.

Grandillo, S., and S. D. Tanksley, 1996 QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium.* Theor. Appl. Genet. **92:** 935–951.

Hospital, F., and C. Chevalet, 1993 Effect of population size and linkage on optimal selection intensity. Theor. Appl. Genet. **86:** 775–780.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136:** 1447–1455.

Kacser, H., and J. A. Burns, 1973 The control of flux. Symp. Soc. Exp. Biol. **27:** 65–104.

Kacser, H., and J. A. Burns, 1981 The molecular basis of dominance. Genetics **97:** 639–666.

Kearsey, M. J., and A. G. L. Farquhar, 1998 QTL analysis in plants; where are we now? Heredity **80:** 137–142.

Kendall, M. G., 1955 *Rank Correlation Methods.* Griffin, London.

Lee, S. H., M. A. Bailey, M. A. R. Mian, T. E. Carter Jr, E. R. Shipe *et al.*, 1996 RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor. Appl. Genet. **93:** 649–657.

Lin, Y. R., K. F. Schertz and A. H. Paterson, 1995 Comparative analysis of QTL affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics **141:** 391–411.

Liu, S.-C., S. P. Kowalski, T.-H. Lan, K. A. Feldmann and A. H. Paterson, 1996 Genome-wide high resolution mapping by recurrent intermating using *Arabidopsis thaliana* as a model. Genetics **142:** 247–258.

Mackay, T. F. C., 1996 The nature of quantitative genetic variation revisited: lessons from Drosophila bristles. Bioessays **18:** 113–121.

Maughan, P. J., M. A. S. Maroof and G. R. Buss, 1996 Molecular-marker analysis of seed-weight: genomic locations, gene action, and evidence for orthologous evolution among three legume species. Theor. Appl. Genet. **93:** 574–579.

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

Orr, H. A., 1999 The evolutionary genetics of adaptation: a simulation study. Genet. Res. **74:** 207–214.

Paterson, A. H., J. W. de Verna, B. Lanini and S. D. Tanksley, 1990 Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecific cross of tomato. Genetics **124:** 735–742.

Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch *et al.*, 1991 Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics **127:** 181–197.

Paterson, A. H., Y. R. Lin, Z. Li, K. F. Schertz, J. F. Doebley *et al.*, 1995 Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science **269:** 1714–1718.

Rodolphe, F., and M. Lefort, 1993 A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics **134:** 1277–1288.

SAS Institute, 1988 *SAS Procedure's Guide*, Release 6.03 Ed. SAS Institute, Cary, NC.

Schön, C. C., A. E. Melchinger, J. Boppenheimer, E. Brunklaus-Jung, R. G. Herrmann *et al.*, 1994 RFLP mapping in maize: quantitative trait loci affecting testcross performance of elite european flint lines. Crop Sci. **34:** 378–388.

Scheffé, H., 1959 *The Analysis of Variance.* Wiley, New York.

Shrimpton, A. E., and A. Robertson, 1988 The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster.* 2. Distribution of third chromosome bristle effects within chromosome sections. Genetics **118:** 445–459.

Sing, C. F., and E. Boerwinkle, 1987 Genetic architecture of inter-individual variability in apolipoprotein, lipoprotein and lipid phenotypes, pp. 99–121 in *Molecular Approaches to Human Polygenic Diseases. CIBA Foundation Symposium No. 130*, edited by G. Bock and G. M. Collins. Wiley, Chichester, UK.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry: The Principles and Practice of Statistics in Biological Research.* W. H. Freeman, New York.

Stam, P., 1991 Some aspects of QTL analysis. Proceedings of the Eighth Meeting of the Eucarpia Section Biometrics in Plant Breeding, July 1991, Brno, Czech Republic.

Timmerman-Vaughan, G. M., J. A. McCallum, T. J. Frew, N. F. Weeden and A. C. Russel, 1996 Linkage mapping of quantitative trait loci controlling seed weight in pea (*Pisum sativum* L.). Theor. Appl. Genet. **93:** 431–439.

Weeks, D. E., and G. M. Lathrop, 1995 Polygenic disease: methods for mapping complex disease traits. Trends Genet. **11:** 513–519.

Zehr, B. E., J. W. Dudley, J. Chojecki, M. A. Saghai Maroof and R. P. Mowers, 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., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: P. D. Keightley

APPENDIX A: **Effect of environmental variation on the percentage of phenotypic variation explained by a QTL, estimated by ANOVA or multiple regression**

We consider here a trait controlled by Q QTL.

One-QTL analysis of variance: The SST of the trait in an F_2 population of N individuals can be decomposed as in Table A1, where κ is the number of genotypic classes at locus q ($\kappa = 3$ in an F_2 population), σ_q^2 is the genetic variance at the QTL q, σ_R^2 is the residual variance, and $\tilde{n} = (N - 1)/(\kappa - 1)$ is a term taking into account the sampling of individuals in the genotypic classes at the QTL q (Charcosset and Gallais 1996). The test statistic of the effect of the locus q is F = mean square $(\mathrm{MS})_q$/residual mean square (MSR), and follows a noncentral Fisher distribution, with $\kappa - 1$ and $N - \kappa$ as degrees of freedom, and a noncentrality parameter ϕ_q (Scheffé 1959):

$$\phi_q = (\kappa - 1)\,\tilde{n}\frac{\sigma_q^2}{\sigma_R^2}.$$

Multiple regression: The SST of the trait in an F_2 population of N individuals can be decomposed as in Table A2, where σ_q^{2*} is the genetic variance at the QTL

TABLE A1

Decomposition of the phenotypic variation with one-QTL analysis of variance

Source of variation	SS	d.f.	E(MS)
QTL $_q$	SS_q	$\kappa - 1$	$\sigma_R^2 + \tilde{n}\,\sigma_q^2$
Residual	SSR	$N - \kappa$	σ_R^2
Total	SST	$N - 1$	σ_P^2

TABLE A2

Decomposition of the phenotypic variation with multiple regression

Source of variation	SS	d.f.	E(MS)
QTL 1	SS_1^*	$\kappa - 1$	$\sigma_R^2 + \tilde{n}\,\sigma_1^{2*}$
⋮	⋮	⋮	⋮
QTL Q	SS_Q^*	$\kappa - 1$	$\sigma_R^2 + \tilde{n}\,\sigma_Q^{2*}$
Residual	SSR*	$N - \kappa - (Q-1)(\kappa - 1)$	σ_R^{2*}
Total	SST	$N - 1$	σ_P^2

q, taking into account the other QTL, $\sigma_R^2{}^*$ is the residual variance, and the other parameters are the same as in Table A1. The test statistic of the effect of the locus q is $F = MS_q^*/MSR^*$, and follows a noncentral Fisher distribution, with $\kappa - 1$ and $N - \kappa - (Q - 1)(\kappa - 1)$ as degrees of freedom, and a noncentrality parameter ϕ_q^*:

$$\phi_q^* = (N-1)\frac{\sigma_q^{2*}}{\sigma_R^{2*}}.$$

As $SST \neq \Sigma_{q=1}^{Q} SS_q^* + SSR^*$, there is no simple relationship between F and the R_q^2 estimated with the multiple regression. However, we can note that Cramer (1987) established the distribution of the global R square for the multiple linear regression with fixed regressors, as a function of the sample size. But as we are interested in partial R^2 it was not possible to use Cramer's results.

APPENDIX B: **Effect of linkage disequilibrium on the percentage of phenotypic variation explained by a QTL, estimated by ANOVA or multiple regression**

We consider here the simple case of a trait controlled by three QTL (Q_1, Q_2, Q_3) being located on the same chromosome. The phenotypic value of the individual i of an F_2 population is

$$z_i = \mathcal{G}_{jkl} + \varepsilon_i, \qquad \text{(B1)}$$

where j, k, l are indices for the number of high alleles (0, 1, or 2) at QTL Q_1, Q_2, Q_3, respectively; $\mathcal{G}_{jkl}$ is the genetic value of individual i; and ε_i is the random environmental deviation. $\mathcal{G}_{jkl}$ is determined according to an additive model,

$$\mathcal{G}_{jkl} = \mathcal{M} + (j-1)a_{Q_1} + (k-1)a_{Q_2} + (l-1)a_{Q_3}, \qquad \text{(B2)}$$

where $\mathcal{M}$ is a constant, and a_{Q1}, a_{Q2}, and a_{Q3} are the allelic additive effects of QTL Q_1, Q_2, and Q_3, respectively.

Let τ_{12}, τ_{23}, τ_{13} be the recombination rates between QTL Q_1 and Q_2, Q_2 and Q_3, and Q_1 and Q_3, respectively. p_{abc} is the frequency of the gamete produced by the F_1 hybrid with allele a, b, and c at QTL Q_1, Q_2, and Q_3, respectively. The eight gametic frequencies are

$$\left.\begin{matrix} p_{000} = p_{111} = E, & p_{001} = p_{110} = F \\ p_{100} = p_{011} = G, & p_{010} = p_{101} = H \end{matrix}\right\} \text{ with } E + F + G + H = \frac{1}{2}.$$

There are four cases for the parental gametic phase (coupling or repulsion) between QTL Q_1 and Q_2, and Q_2 and Q_3 (Table B1), and the corresponding gametic frequencies E, F, G, and H are given in Table B2.

In the F_2 population, the linkage disequilibria between Q_1 and Q_2 (D_{12}), Q_2 and Q_3 (D_{23}), and Q_1 and Q_3 (D_{13}) are

$$D_{12} = E + F - \frac{1}{4} = \frac{1}{2}(E + F - G - H)$$

$$D_{23} = E + G - \frac{1}{4} = \frac{1}{2}(E + G - F - H)$$

$$D_{13} = E + H - \frac{1}{4} = \frac{1}{2}(E + H - F - G).$$

It is worth noting that

$$D_{13} = 4D_{12}\,D_{23}.$$

The genotype frequencies (f_{jkl}) can be deduced from gametic frequencies. For example, $f_{011} = 2p_{000}p_{011} + 2p_{010}p_{001} = 2EF + 2GH$.

In this model, the mean phenotypic value of the population is

TABLE B1

Possible parental gametic phases with three QTL

$Q_1 - Q_2$	$Q_2 - Q_3$	Parental genotypes
Coupling	Coupling	$\frac{000}{000}, \frac{111}{111}$
Repulsion	Repulsion	$\frac{010}{010}, \frac{101}{101}$
Repulsion	Coupling	$\frac{011}{011}, \frac{100}{100}$
Coupling	Repulsion	$\frac{001}{001}, \frac{110}{110}$

TABLE B2

Gametic frequencies at three QTL in a F$_2$ population

$Q_1 - Q_2/Q_2 - Q_3$	E	F	G	H
Coupling/coupling	$\frac{(1-\tau_{12})(1-\tau_{23})}{2}$	$\frac{(1-\tau_{12})\ \tau_{23}}{2}$	$\frac{\tau_{12}\ (1-\tau_{23})}{2}$	$\frac{\tau_{12}\ \tau_{23}}{2}$
Repulsion/repulsion	$\frac{\tau_{12}\ \tau_{23}}{2}$	$\frac{\tau_{12}\ (1-\tau_{23})}{2}$	$\frac{(1-\tau_{12})\ \tau_{23}}{2}$	$\frac{(1-\tau_{12})(1-\tau_{23})}{2}$
Repulsion/coupling	$\frac{\tau_{12}\ (1-\tau_{23})}{2}$	$\frac{\tau_{12}\ \tau_{23}}{2}$	$\frac{(1-\tau_{12})(1-\tau_{23})}{2}$	$\frac{(1-\tau_{12})\ \tau_{23}}{2}$
Coupling/repulsion	$\frac{(1-\tau_{12})\ \tau_{23}}{2}$	$\frac{(1-\tau_{12})(1-\tau_{23})}{2}$	$\frac{\tau_{12}\ \tau_{23}}{2}$	$\frac{\tau_{12}\ (1-\tau_{23})}{2}$

$$\mu = \sum_j\sum_k\sum_l f_{jkl}\mathcal{G}_{jkl} = \mathcal{M}$$

and the total genetic variance is

$$\begin{aligned}\sigma^2_{\mathrm{G}} &= \sum_j\sum_k\sum_l f_{jkl}\ (\mathcal{G}_{jkl} - \mu)^2\\ &= \frac{1}{2}(a^2_{Q_1} + a^2_{Q_2} + a^2_{Q_3}) + 4D_{12}\ a_{Q_1}\ a_{Q_2}\\ &\quad + 4D_{23}\ a_{Q_2}\ a_{Q_3} + 4D_{13}\ a_{Q_1}\ a_{Q_3}. \qquad \text{(B3)}\end{aligned}$$

One-QTL analysis of variance: With one-QTL ANOVA, the statistical model for the phenotypic value of an individual *i*, with the genotype *j* at the QTL considered, is

$$z_{ij} = G_j + g_{ij} + \varepsilon_{ij}, \qquad \text{(B4)}$$

where G_j is the effect of genotype *j* at the QTL, g_{ij} is a genetic component due to segregation at the two other QTL, and ε_{ij} is the random environmental deviation. The mean phenotypic value for genotype *j* at the QTL is

$$\mu_j = E_i(z_{ij}) = G_j + E_i(g_{ij}),$$

which is the conditional expectation for the phenotype, given the genotype at the QTL. We show below that, in this simple case, $E_i(g_{ij}) \neq 0$ if there is linkage between the QTL considered and the other ones.

For example, for the QTL Q_2, the mean phenotypic value depends on genotype frequencies f_{jkl} at the three QTL, as well as on genetic values $\mathcal{G}_{jkl}$,

$$\mu_{Q_{2_k}} = \sum_{j=0}^{2}\sum_{l=0}^{2}\frac{f_{jkl}}{f_k}\mathcal{G}_{jkl}\,,$$

and the variance contributed by QTL Q_2 is

$$\sigma^2_{Q_2} = \sum_{k=0}^{2} f_k(\mu_{Q_{2_k}} - \mu)^2.$$

In the infinite F$_2$ population that we consider, there are three different genotypes at one QTL, with genotypic frequencies $f_0 = 1/4$, $f_1 = 1/2$, and $f_2 = 1/4$. As an example, for the QTL Q_2, we then have

$$\begin{aligned}\mu_{Q_{2_0}} &= \mathcal{M} - a_{Q_2} - 4D_{12}\ a_{Q_1} - 4D_{23}\ a_{Q_3}\\ \mu_{Q_{2_1}} &= \mathcal{M}\\ \mu_{Q_{2_2}} &= \mathcal{M} + a_{Q_2} + 4D_{12}\,a_{Q_1} + 4D_{23}\ a_{Q_3}.\end{aligned}$$

Thus, we can calculate the contribution of each QTL:

$$\sigma^2_{Q_1} = \frac{1}{2}(a_{Q_1} + 4D_{12}\ a_{Q_2} + 4D_{13}\,a_{Q_3})^2 \qquad \text{(B5)}$$

$$\sigma^2_{Q_2} = \frac{1}{2}(a_{Q_2} + 4D_{12}\ a_{Q_1} + 4D_{13}\,a_{Q_3})^2 \qquad \text{(B6)}$$

$$\sigma^2_{Q_3} = \frac{1}{2}(a_{Q_3} + 4D_{23}\ a_{Q_2} + 4D_{13}\,a_{Q_1})^2. \qquad \text{(B7)}$$

Hence, the sum of the individual contributions of each QTL is not equal to the total genetic variance (B3) when QTL are linked ($D \neq 0$).

Multiple regression: With multiple regression, we take into account all the QTL, and the statistical model for the phenotypic value of an individual *i*, with the genotypes *j*, *k*, and *l* at the QTL Q_1, Q_2, and Q_3, respectively, is

$$z_{ijkl} = G_j + G_k + G_l + \varepsilon_{ijkl}\,, \qquad \text{(B8)}$$

where ε_{ijkl} is the random environmental deviation. The genetic variances contributed by each QTL are computed conditionally on the other QTL declared in the model,

$$\sigma^2_{Q_1} = \sigma^2_{\mathrm{G}} - \sigma^2_{Q_{23}},\quad \sigma^2_{Q_2} = \sigma^2_{\mathrm{G}} - \sigma^2_{Q_{13}},\quad \sigma^2_{Q_3} = \sigma^2_{\mathrm{G}} - \sigma^2_{Q_{12}}\,, \qquad \text{(B9)}$$

where

$$\begin{aligned}\sigma^2_{Q_{12}} &= \sum_{j=0}^{2}\sum_{k=0}^{2} f_{jk}\ (\mu_{jk} - \mu)^2\\ \sigma^2_{Q_{13}} &= \sum_{j=0}^{2}\sum_{l=0}^{2} f_{jl}\ (\mu_{jl} - \mu)^2\\ \sigma^2_{Q_{23}} &= \sum_{k=0}^{2}\sum_{l=0}^{2} f_{kl}\ (\mu_{kl} - \mu)^2\end{aligned}$$

are the genetic variances contributed by the other QTL, and

$$\mu_{jk} = \sum_{l=0}^{2} \frac{f_{jkl}}{f_{jk}} \mathcal{G}_{jkl}, \quad \mu_{jk} = \sum_{k=0}^{2} \frac{f_{jkl}}{f_{jl}} \mathcal{G}_{jkl}, \quad \mu_{kl} = \sum_{j=0}^{2} \frac{f_{jkl}}{f_{kl}} \mathcal{G}_{jkl}$$

are, respectively, the mean phenotypic values for genotype *jk* at QTL Q_1 and Q_2, *jl* at QTL Q_1 and Q_3, and *kl* at QTL Q_2 and Q_3. Using the genetic model defined in (B2), we find

$$\sigma^2_{Q_{12}} = \sigma^2_G - \frac{1}{2} a^2_{Q_3} \left[\frac{1 - 16\,(D^2_{12} + D^2_{13} + D^2_{23}) + 128\, D_{12}\, D_{13}\, D_{23}}{1 - 16\, D^2_{12}}\right]$$

$$\sigma^2_{Q_{13}} = \sigma^2_G - \frac{1}{2} a^2_{Q_2} \left[\frac{1 - 16\,(D^2_{12} + D^2_{13} + D^2_{23}) + 128\, D_{12}\, D_{13}\, D_{23}}{1 - 16\, D^2_{13}}\right]$$

$$\sigma^2_{Q_{23}} = \sigma^2_G - \frac{1}{2} a^2_{Q_1} \left[\frac{1 - 16\,(D^2_{12} + D^2_{13} + D^2_{23}) + 128\, D_{12}\, D_{13}\, D_{23}}{1 - 16\, D^2_{23}}\right].$$

Thus, following (B9), the variances contributed by each QTL are

$$\sigma^2_{Q_1} = \frac{1}{2} a^2_{Q_1}\,(1 - 16D^2_{12}) \qquad \text{(B10)}$$

$$\sigma^2_{Q_2} = \frac{1}{2} a^2_{Q_2} \left[\frac{(1 - 16D^2_{12})\,(1 - 16D^2_{23})}{1 - 256D^2_{12}\, D^2_{23}}\right] \qquad \text{(B11)}$$

$$\sigma^2_{Q_3} = \frac{1}{2} a^2_{Q_3}\,(1 - 16D^2_{23}), \qquad \text{(B12)}$$

and the sum of the individual contributions of each QTL is

$$\frac{1}{2}(a^2_{Q_1} + a^2_{Q_2} + a^2_{Q_3}) - 8\left[a^2_{Q_1}\, D^2_{12} + a^2_{Q_2} \frac{D^2_{12} + D^2_{23} - 32\, D^2_{12}\, D^2_{23}}{1 - 256D^2_{12}\, D^2_{23}} + a^2_{Q_3}\, D^2_{23}\right]$$

and is not equal to the total genetic variance (B3) when QTL are linked ($D \neq 0$).

Note that, in an F_2 population, with multiple regression, the effect of a QTL q does not involve the effects of the QTL that are linked to q, but only the linkage disequilibria between these QTL and QTL q.