Quantitative Trait Loci: a meta-analysis

Bruno Goffinet, Sophie Gerber

Quantitative Trait Loci: A Meta-analysis

Bruno Goffinet* and Sophie Gerber†,1

**INRA, Laboratoire de Biométrie et d'Intelligence Artificielle, F-31326 Castanet-Tolosan Cedex, France and †INP/ENSAT, Laboratoire de Biotechnologie et d'Amélioration des Plantes, F-31326 Castanet-Tolosan Cedex, France*

Manuscript received August 20, 1999
Accepted for publication January 24, 2000

ABSTRACT

This article presents a method to combine QTL results from different independent analyses. This method provides a modified Akaike criterion that can be used to decide how many QTL are actually represented by the QTL detected in different experiments. This criterion is computed to choose between models with one, two, three, etc., QTL. Simulations are carried out to investigate the quality of the model obtained with this method in various situations. It appears that the method allows the length of the confidence interval of QTL location to be consistently reduced when there are only very few "actual" QTL locations. An application of the method is given using data from the maize database available online at http://www.agron.missouri.edu/.

A meta-analysis consists of combining data from different sources in a single study. This technique is mainly used by researchers in medical, social, and behavioral sciences (Hedges and Olkin 1985). Its application in genetics and evolution is illustrated in recent publications (Britten 1996; Allison and Heo 1998; Van Zandt and Mopper 1998), and the usefulness of these kinds of methods for pooling information when raw data are not available is emphasized (Lander and Kruglyak 1995; Allison and Heo 1998).

Since the first publication of a quantitative trait locus (QTL) localization using molecular markers (Paterson *et al.* 1988) a large number of species have been studied for numerous markers and traits. Some of the data obtained are now available online, as in the maize database (at http://www.agron.missouri.edu), where structured QTL data sets are continuously updated. Having data concerning different populations, it would be interesting to know whether QTL identified for a given trait in one population correspond to those detected in other populations, or whether QTL locations identified in one species correspond to QTL or other types of loci detected in corresponding regions of other species. The QTL aspect of the database was created to encourage systematic description of QTL studies and to facilitate these kinds of comparisons (Byrne *et al.* 1995). Hence, a recent revision of the QTL described in the database suggests that QTL associations coincide with clusters of qualitative developmental genes (Khavkin and Coe 1997, 1998).

Comparative analysis of QTL between species reveals the existence of homologous QTL for plant height and maturity within the Poaceae (sorghum, maize, rice, wheat, and barley; Lin *et al.* 1995). Similar observations for traits involved in domestication suggest that few genes with a large effect have determined the phenotypes studied (Paterson *et al.* 1995). Comparing species is also a means to find new QTL, increasing their potential use for plant breeding, as in the tomato (Fulton *et al.* 1997). Moreover, the existence of small common regions on linkage maps between taxa that diverged a long time ago may provide the opportunity to extend results obtained in one species (Paterson *et al.* 1996) and permit the cross-utilization of resources that have been developed for a given species (Kowalski *et al.* 1994).

Several statistical methods to detect QTL have been developed (Jansen 1996). A QTL, once detected, is described by its position on a linkage group, and possibly a confidence interval around this position, an R2 or a lod score. When several crosses are available and studied simultaneously for the same trait, a first statement is to consider that the QTL are common to both crosses but that their alleles are different. Detecting QTL by interval mapping consists then in testing at every point of the chromosome the existence of a QTL with a potential effect on each cross. This is the case when offspring of different males are considered in the same study in animal genetics with a model where the QTL effect is different from one male to the other. This technique is a way of increasing the QTL detection power when the positions of the QTL are effectively the same through the different crosses. However, when these positions are different, this method can mix them and

Corresponding author: Bruno Goffinet, INRA, Laboratoire de Biométrie et d'Intelligence Artificielle, Centre de Recherches de Toulouse, F-31326 Castanet-Tolosan Cedex, France.
E-mail: goffinet@toulouse.inra.fr

[1]*Present address:* INRA, Recherches Forestières, Laboratoire de Génétique et Amélioration des Arbres Forestiers, BP 45, F-33611 Gazinet Cedex, France.

suggest "consensus" positions that do not actually correspond to any real position. The problem lies in testing between one or several QTL in the same linkage group (Hyne and Kearsey 1995; Goffinet and Mangin 1998). Another statement is to estimate QTL independently in every cross and to test whether they are located on the same place or not. When two crosses are considered, the similarity of the positions detected in the crosses can be tested by using confidence intervals for each position. When more than two QTL are involved, the problem is more acute. The question to be answered is not only whether there is a common position but also to determine the number of different positions.

In this article we suggest an approach for choosing the best model to fit a set of data. Our aim was to elaborate a meta-analysis of several QTL related to the same trait and mapped on the same linkage group in different independent studies. The question we wanted to address was the following: How many "real" QTL do the QTL detected in the different studies represent—one, two, three, or as many as the number detected throughout the studies? Once this question is answered, the positions of the real QTL can be estimated. This approach should help to gather data obtained from different populations and extract meaningful results for the species under investigation.

METHODOLOGY

The QTL experiment summary: Consider a set of n QTL experiments concerning the same linkage group. These different experiments may represent several crosses between different lines, or several sires, or different traits, or different locations for the same trait, or different environmental conditions and experimental designs.

We consider that for each experiment i ($i = 1, \ldots, n$), the summary of the information is the estimated position of the QTL $\hat{x}_i$ for this experiment in this linkage group. We assume that the $\hat{x}_i$ are normally distributed around the true position x_i of the QTL in experiment i, with a variance $\text{var}(\hat{x}_i) = \gamma_{E,i}$. As this variance can be generally estimated with a large number of observations, we assume that it is consistently estimated and therefore can be considered as known. Nevertheless, we investigate in the simulation section the effect of an imperfect estimation of this parameter.

This Gaussian and unbiased approximation can be considered as correct for QTL with a large effect. In these cases, one can use the classical asymptotic Gaussian distribution of the maximum-likelihood estimation of the parameters. For QTL with small effects, Mangin *et al.* (1994) have shown that it is not perfectly correct. Nevertheless, it is a simple and useful approximation and we consider it as correct for all the detected QTL.

The other information available in a QTL experiment is the estimated QTL effect. We do not make use of this information in this article, except as a possible way, when combined with map density and the number of observations, of estimating the variance of the QTL's estimated position. Actually, the larger the QTL effect, the smaller the $\text{var}(\hat{x}_i)$.

The n experiments are considered as independent. This is clearly correct when the individuals measured in the different experiments are different. It is an approximation when these experiments represent different traits measured on the same individuals or when two or more QTL are detected for the same trait in an experiment. We study in the simulation section the effect of considering independence between the experiments when there are actually dependences between some experiments. Independence between experiment i and i' means independence between $\hat{x}_i$ and $\hat{x}_{i'}$; that is, basically the individuals used in the two experiments are not the same, even supposing that the parent lines are the same.

The set of $\hat{x}_i$, $i = 1, n$ is denoted $\hat{X}$.

The different models: Let $k = 1, \ldots, n$ represent different models for the real position x_i of the n QTL. In model $k = 1$, we consider that all the n QTL are located at a single position. In model k, we consider that there are k different positions for the n QTL, and model n corresponds to the case where the n QTL are located at n different positions.

For each experiment i and model k we denote $\tilde{x}_i^{[k]}$ the estimate of position x_i. We use the following estimates:

$k = 1$: $\tilde{x}_i^{[1]} = \bar{x} = {}^1\!/_n \Sigma_i \hat{x}_i$.

$k = 2$: $\mu_1^{[2]}$ and $\mu_2^{[2]}$ are the maximum-likelihood estimates of the possible values of x_i in the two-population mixture model. To estimate these parameters, we consider all the possible distributions of the n QTL into two groups. For each distribution, we compute the maximum-likelihood estimator of the mean of each group and choose the best distribution as the distribution maximizing the likelihood.

We have $\tilde{x}_i^{[2]} = \mu_1^{[2]}$ if $|\hat{x}_i - \mu_1^{[2]}| < |\hat{x}_i - \mu_2^{[2]}|$ and $\tilde{x}_i^{[2]} = \mu_2^{[2]}$ otherwise.

$k < n$: The same rule as for $k = 2$ applies with $\mu_1^{[k]}$, $\mu_2^{[k]}, \ldots, \mu_k^{[k]}$ the k possible values of the k-population mixture model. Notation $[k]$ represents the number of QTL in this model. As for $k = 2$, we consider all the possible distributions of the n QTL into k groups and choose the distribution with the maximum likelihood.

We have $\tilde{x}_i^{[k]} = \mu_j^{[k]}$, where j is such that $|\hat{x}_i - \mu_j^{[k]}|$ is minimum for $j = 1, \ldots, k$.

$k = n$: $\tilde{x}_i^{[n]} = \hat{x}_i$.

Model selection: The problem lies in finding a crite-

rion to choose from the different models $k = 1, \ldots, n$. It is known (Titterington *et al.* 1985) that the Akaike criterion is not a correct way of comparing models in the case of mixture models. We propose herein an adaptation of the Akaike criterion to deal with our models.

Consider a model with k parameters $\Theta^{[k]} = \mu_1^{[k]}, \mu_2^{[k]}, \ldots, \mu_k^{[k]}$ and the n corresponding values of the QTL positions $X = (x_i)_{i=1,n}$. The log-likelihood of the observed vector $\hat{X}$ is denoted $L(\Theta^{[k]}, X; \hat{X})$. We denote k_0 the actual number of parameters, $\Theta_0^{[k]}$ the actual value of the parameters, and X_0 the actual value of the n QTL positions. The maximum-likelihood estimates are denoted $\hat{\Theta}^{[k]}$ and $\tilde{X}^{[k]} = (\tilde{x}_i^{[k]})_{i=1,n}$ and the corresponding log-likelihood $L(\hat{\Theta}^{[k]}, \tilde{X}^{[k]}; \hat{X})$.

The aim of the Akaike criterion (Sakamoto *et al.* 1986) is to estimate the mean expected log-likelihood (MELL):

$$\text{MELL} = E_{(\tilde{\Theta}^{[k]}, \tilde{X}^{[k]})}(E_{\hat{X}^*}(L(\hat{\Theta}^{[k]}, \tilde{X}^{[k]}; \hat{X}^*))).$$

In the second expectation $E_{\hat{X}^*}$, the estimated values $\hat{\Theta}^{[k]}$, $\tilde{X}^{[k]}$ are fixed, and the expectation is taken for independent possible values of observations $\hat{X}^*$, with the same probability distribution function as $\hat{X}$.

In regular situations, it is well known that $L(\hat{\Theta}^{[k]}, \tilde{X}^{[k]}; \hat{X}) - k$, where k is the number of free parameters of the model, is an asymptotically unbiased estimator of MELL. Therefore, it is recommended to choose a model minimizing the Akaike information criterion, $\text{AIC} = -2 \times L(\hat{\Theta}^{[k]}, \tilde{X}^{[k]}; \hat{X}) + 2 \times k$.

In our cases of mixtures models, $L(\hat{\Theta}^{[k]}, \tilde{X}^{[k]}; \hat{X}) - k$ is not an unbiased estimator of MELL, except for $k = 1$ and $k = n$. We propose to estimate numerically $\text{bias}(X_0, k_0; k) = \text{MELL} - E(L(\hat{\Theta}^{[k]}, \tilde{X}^{[k]}; \hat{X}))$ in different situations. Table 1 shows the values of this bias for different values of n and k, and different values of k_0 and X_0 of the actual model, that is, when using a model with k parameters when there are actually k_0 parameters and the actual parameter values are X_0. The different configurations l, $l = 1, 10$ are described in Table 2. The computations are based on $\sigma_{E,i} = \sqrt{\gamma_{E,i}} = 10$ cM.

Consider first, for example, the case $n = 20$, where more configurations are studied. The value of bias depends strongly upon the value of the number k_0 of parameters in the actual model. Nevertheless, the main aim of correcting the log-likelihood is to prevent the choice of a model with more than k_0 parameters when the number of parameters is actually k_0. We observed that the difference between $\text{bias}(X_0, k_0; k_0)$ and $\text{bias}(X_0, k_0; k_0 + 1)$ depends little upon the values of X_0. For example, when the actual model has $k_0 = 2$ parameters, the difference between the bias when using a model with 3 parameters and a model with 2 parameters is (see Table 1) $12.7 - 3.9 = 8.8$ [configuration (config) 2], or $10.6 - 2.0 = 8.6$ (config 3), or $10.3 - 2.1 = 8.1$ (config 4). This value tends to converge to a stable value as the difference between the parameters μ_i is increasing. We may therefore use the value 8.1 as a limit value. These limit values are 13 for $k_0 = 1$, 8.1 for $k_0 = 2$, 6.7 for $k_0 = 3$, and 15.9 for $k_0 = 4$. In this case, $k_0 + 1$ is taken as n. We observed slightly different limit values when using unbalanced configurations (data not shown).

We propose, therefore, to use the following expressions of AIC*(k) to choose from the models $k = 1, 2, 3, 4, n = 20$:

$$\text{AIC}^*(1) = -2 \times (L(\hat{\Theta}^{[1]}, \tilde{X}^{[1]}; \hat{X}) - 1)$$

$$\text{AIC}^*(2) = -2 \times (L(\hat{\Theta}^{[2]}, \tilde{X}^{[2]}; \hat{X}) - 1 - 13)$$

$$\text{AIC}^*(3) = -2 \times (L(\hat{\Theta}^{[3]}, \tilde{X}^{[3]}; \hat{X}) - 1 - 13 - 8.1)$$

$$\text{AIC}^*(4) = -2 \times (L(\hat{\Theta}^{[4]}, \tilde{X}^{[4]}; \hat{X}) - 1 - 13 - 8.1 - 6.7)$$

$$\text{AIC}^*(20) = -2 \times (L(\hat{\Theta}^{[n]}, \tilde{X}^{[n]}; \hat{X}) - 1 - 13 - 8.1 - 6.7 - 15.9)$$

It appears that these coefficients are approximately constant or a linear function of n when n changes. We can therefore propose the following expressions for AIC*(k)

TABLE 1

Value of bias(X_0,k_0;k) in different situations

	n:	10					20					40				
	k:	1	2	3	4	n	1	2	3	4	n	1	2	3	4	n
$k_0 = 1$	Config 1	1	7.6	9.2	9.7	10	1	14.0	17.5	18.8	20	1	27.1	33.8	36.6	40
$k_0 = 2$	Config 2	1	2.9	7.4	8.9	10	1	3.9	12.7	16.3	20	1	5.3	23.2	30.5	40
	Config 3	1	2.1	6.7	8.6	10	1	2.0	10.6	15.4	20	1	1.9	18.4	28.3	40
	Config 4						1	2.1	10.3	15.2	20					
$k_0 = 3$	Config 5	1	9.1	4.4	7.5	10	1	16.3	5.4	12.3	20	1	29.7	7.5	21.2	40
	Config 6	1	1.9	3.8	7.2	10	1	2.5	4.4	11.2	20	1	0.5	5.2	17.9	40
	Config 7						1	27.6	3.0	9.7	20					
$k_0 = 4$	Config 8	1	4.2	9.8	5.5	10	1	4.6	16.6	7.1	20	1	6.6	32.0	10.0	40
	Config 9	1	2.6	10.5	4.3	10	1	2.4	16.5	4.6	20	1	2.9	25.8	4.9	40
	Config 10						1	2.3	26.0	4.1	20					

TABLE 2

Description of the configurations for $n = 10$ experiments

Configurations	m/type	μ_i	$\sigma_{E,i}$	$\sigma_{E,i;0}$	Configurations	m/type	μ_i	$\sigma_{E,i}$	$\sigma_{E,i;0}$
Config 1	10	0.0	0.1	0.1					
Config 2	5	0.0	0.1	0.1	Config 14	2	0.0	0.1	0.1
	5	0.5	0.1	0.1		3	0.4	0.1	0.1
						2	0.7	0.1	0.1
						3	1.1	0.1	0.1
Config 3	5	0.0	0.1	0.1	Config 15	1	0.0	0.1	0.1
	5	1.0	0.1	0.1		1	0.1	0.1	0.1
						1	0.15	0.1	0.1
						1	0.33	0.1	0.1
						1	1.48	0.1	0.1
						1	0.55	0.1	0.1
						1	0.58	0.1	0.1
						1	0.67	0.1	0.1
						1	0.68	0.1	0.1
						1	0.85	0.1	0.1
Config 4	5	0.0	0.1	0.1	Config 16	5	0.0	0.05	0.05
	5	2.0	0.1	0.1		5	0.4	0.05	0.05
Config 5	4	0.0	0.1	0.1	Config 17	5	0.0	0.07	0.07
	3	0.5	0.1	0.1		5	0.4	0.07	0.07
	3	1.0	0.1	0.1					
Config 6	3	0.0	0.1	0.1	Config 18	5	0.0	0.15	0.15
	4	1.0	0.1	0.1		5	0.4	0.15	0.15
	3	1.5	0.1	0.1					
Config 7	3	0.0	0.1	0.1	Config 19	5	0.0	0.2	0.2
	4	2.5	0.1	0.1		5	0.4	0.2	0.2
	3	4.0	0.1	0.1					
Config 8	3	0.0	0.1	0.1	Config 20	5	0.0	0.05	0.05
	2	0.5	0.1	0.1		5	0.8	0.05	0.05
	2	1.0	0.1	0.1					
	3	1.5	0.1	0.1					
Config 9	2	0.0	0.1	0.1	Config 21	5	0.0	0.07	0.07
	3	0.7	0.1	0.1		5	0.8	0.07	0.07
	3	1.4	0.1	0.1					
	2	2.0	0.1	0.1					
Config 10	2	0.0	0.1	0.1	Config 22	5	0.0	0.15	0.15
	3	1.5	0.1	0.1		5	0.8	0.15	0.15
	3	3.0	0.1	0.1					
	2	4.5	0.1	0.1					
Config 11	5	0.0	0.1	0.1	Config 23	5	0.0	0.2	0.2
	5	0.4	0.1	0.1		5	0.8	0.2	0.2
Config 12	5	0.0	0.1	0.1					
	5	0.8	0.1	0.1					
Config 13	3	0.0	0.1	0.1					
	4	0.4	0.1	0.1					
	3	0.7	0.1	0.1					
Config 24	5	0.0	0.1	0.1	Config 28	5	0.0	0.1	0.1
	3	0.0	0.15	0.15		3	0.0	0.1	0.15
	2	0.0	0.07	0.07		2	0.0	0.1	0.07

(*continued*)

TABLE 2

(Continued)

Configurations	m/type	μ_i	$\sigma_{E,i}$	$\sigma_{E,i;0}$	Configurations	m/type	μ_i	$\sigma_{E,i}$	$\sigma_{E,i;0}$
Config 25	2	0.0	0.1	0.1	Config 29	2	0.0	0.1	0.1
	3	0.0	0.07	0.07		3	0.0	0.1	0.07
	3	0.4	0.15	0.15		3	0.4	0.1	0.15
	2	0.4	0.1	0.1		2	0.4	0.1	0.1
Config 26	3	0.0	0.1	0.1	Config 30	3	0.0	0.1	0.1
	2	0.0	0.07	0.07		2	0.0	0.1	0.07
	3	0.8	0.15	0.15		3	0.8	0.1	0.15
	2	0.8	0.1	0.1		2	0.8	0.1	0.1
Config 27	2	0.0	0.1	0.1	Config 31	2	0.0	0.1	0.1
	1	0.0	0.07	0.07		1	0.0	0.1	0.07
	1	0.5	0.1	0.1		1	0.5	0.1	0.1
	1	0.5	0.07	0.07		1	0.5	0.1	0.07
	1	1.0	0.15	0.15		1	1.0	0.1	0.15
	1	1.0	0.1	0.1		1	1.0	0.1	0.1
	2	1.5	0.15	0.15		2	1.5	0.1	0.15
	1	1.5	0.1	0.1		1	1.5	0.1	0.1
Config 32	5	0.0	0.1	0.1	Correlation				
	3	0.0	0.1	0.15	$\rho = 0.8$				
	2	0.0	0.1	0.07	between	1,2; 3,4; etc.			
Config 33	2	0.0	0.1	0.1	Correlation				
	3	0.0	0.1	0.07	$\rho = 0.8$				
	3	0.4	0.1	0.15	between	1,2; 3,4; etc.			
	2	0.4	0.1	0.1					
Config 34	3	0.0	0.1	0.1	Correlation				
	2	0.0	0.1	0.07	$\rho = 0.8$				
	3	0.8	0.1	0.15	between	1,2; 3,4; etc.			
	2	0.8	0.1	0.1					
Config 35	2	0.0	0.1	0.1	Correlation				
	1	0.0	0.1	0.07	$\rho = 0.8$				
	1	0.5	0.1	0.1	between	1,2; 3,4; etc.			
	1	0.5	0.1	0.07					
	1	1.0	0.1	0.15					
	1	1.0	0.1	0.1					
	2	1.5	0.1	0.15					
	1	1.5	0.1	0.1					

For $n = 20$, each configuration is doubled ($\times 4$ for $n = 40$). m/type is the number of experiments of the type described in the line that is with an actual expectation μ_i and actual standard deviation $\sigma_{E,i;0}$; $\sigma_{E,i}$ is the standard deviation used in the simulations for this type.

that can be used for any value of n such that $10 \leq n \leq 40$:

$$\text{AIC}^*(1) = -2 \times (L(\hat{\Theta}^{[1]}, \tilde{X}^{[1]}; \hat{X}) - 1)$$

$$\text{AIC}^*(2) = -2 \times (L(\hat{\Theta}^{[2]}, \tilde{X}^{[2]}; \hat{X}) - 0.7 \times n)$$

$$\text{AIC}^*(3) = -2 \times (L(\hat{\Theta}^{[3]}, \tilde{X}^{[3]}; \hat{X}) - 1.11 \times n)$$

$$\text{AIC}^*(4) = -2 \times (L(\hat{\Theta}^{[4]}, \tilde{X}^{[4]}; \hat{X}) - 1.44 \times n)$$

$$\text{AIC}^*(n) = -2 \times (L(\hat{\Theta}^{[n]}, \tilde{X}^{[n]}; \hat{X}) - 2.27 \times n).$$

Note that we do not propose expressions for $k = 5, \ldots, n - 1$. The reason for that is the inefficiency of the use of the corresponding models when $10 \leq n \leq 40$ and the length of chromosome is shorter than 2 M. Nevertheless, models with $> k = 4$ parameters could be efficient when n becomes >40 or for chromosome length >2 M.

The expressions for AIC*(k) were obtained using a particular situation for the $\sigma_{E,i}$ and independence between the $\hat{x}_i$. Nevertheless, we propose to use these expressions in general situations including different and variable values for the $\sigma_{E,i}$ and nonindependence. Their efficiencies in these situations are investigated by simulations in the following section.

COMPARISON OF MODEL SELECTION STRATEGIES

Alternative strategies and comparison indicators: We now compare the quality of different estimates of x_i

obtained with the two alternative strategies of choosing a model:

strategy S_1. $\tilde{x}_i(S_1) = \hat{x}_i$. This is the "conventional" strategy, which retains the estimated position.

strategy S_2. Choose the model l_2 giving the minimum value of the AIC*(l_2) criterion. The corresponding estimate of x_i is $\tilde{x}_i(S_2) = \tilde{x}_i^{[l_2]}$.

For each of these $h = 1, 2$ strategies, we compute two kinds of indicators:

The mean squared error of prediction $R^{S_h} = {}^1\!/_n \sum_{i=1}^{n} E(x_i - \tilde{x}_i(S_h))^2$.

The length of the confidence interval at 95 and 90% for the position of the QTL. To obtain this length, we compute the quantities $|x_i - \tilde{x}_i(S_h)|$ and calculate the quantiles $q(0.95)$ and $q(0.90)$ of its empirical distribution over all the QTL. The smaller this confidence interval, the better the location estimator $\tilde{x}_i(S_h)$.

Simulation results: We compare different configurations concerning k_0 and X_0 in four steps. In the first step, we consider the standard deviation $\sigma_{E,i} = \sqrt{\gamma_{E,i}}$ as constant among $i = 1, n$ and known; that is, the actual standard deviation $\sigma_{E,i;0}$ used in the simulations is the same as the standard deviation $\sigma_{E,i}$ used in the model. In the second step, the standard deviations are known but different from one observation i to another. In the third step, the standard deviations are different and unknown; that is, the standard deviation $\sigma_{E,i;0}$ used in the simulations is different from the standard deviation of the model. In the fourth step, we investigate the effect of nonindependence between the experiments by adding into the simulation model a correlation $\rho = 0.8$ between $\hat{x}_i$ and $\hat{x}_{i'}$ for $i = 1$ and $i' = 2$, $i = 3$, and $i' = 4$ and so on. This choice is arbitrary. In all these cases, the number of observations is $n = 20$, the $\hat{x}_i$ values are simulated as normally distributed $N(x_i, \gamma_{E,i;0})$, and we perform 500 simulations. The configurations are described in Table 2 and the results in Table 3. The reason for the choice of $n = 20$ is that it is a common number of experiments that are presently found in the literature. The choice of the configurations is linked to the length of maize chromosomes (between 1 and 2 M). The configurations try to cover the range of possible repartitions of QTL positions. It does not try to be a "sample" of the reality as we do not know what the reality is. In Table 3, we give the value of the mean squared error of prediction R^{S_h}, and the mean length of the 90 and 95% confidence interval of the QTL position, for both strategies S_1 and S_2.

Step 1. Configurations 1–23: It appears that the gain obtained with strategy S_2 is substantial in several situations for the different comparison indicators. For example, the length of the 95% confidence interval is divided by 4.5 when using S_2 when there is actually only one QTL position. In several situations, this length is halved. Note that to halve a confidence interval in a QTL experiment, one needs to use four times the initial number of observations. The conventional strategy S_1 becomes equal or better when there are many actual positions (config 15) or when the actual QTL positions are narrow in regard to variance (config 4, 18, and 19). Nevertheless, the greatest loss is ~20% for the confidence intervals. Except for config 13, the conclusions are the same for the three comparison indicators.

Step 2. Configurations 24–27: When comparing config 24 with config 1, and config 26 with config 12, it appears that the variability among the $\sigma_{E,i}$ does not change the behavior of the strategies for all the criteria. Nevertheless, the comparisons between config 25 with config 11 and config 27 with config 8 show that the gain in using S_2 is less when there is a variability among the variances when using the 0.95% confidence interval criterion. The difference between config 25 and 11 is more important than the difference between config 24 and 1 because it is possible to detect two populations whose means differ by 0.4 with $\sigma_{E,i} = 0.1$, but it becomes more difficult when $\sigma_{E,i} = 0.15$.

Step 3. Configurations 28–31: As previously, the comparisons between config 24 and 28 and between config 26 and 30 show some decrease in the gain when using S_2, but not a very substantial one. The gain in using S_2 for the 95% confidence interval continues to decrease when comparing config 25 with 29 and config 27 with 31.

Step 4. Configurations 32–35: Globally the comparisons between config 28 and 32, config 29 and 33, config 30 and 34, and config 31 and 35 show a small decrease in the gain when using S_2 for the different indicators.

Nevertheless, the use of S_2 in all these configurations continues to be advantageous (config 35) or very advantageous (config 32 and 34) for all the indicators. The conclusions are less clear for config 33, as it depends on the indicator.

We do not give the loss in gain for all types of configurations through the three last steps. For example, the series config 20, 26, 30, and 34 have the same behavior as the same kind of series beginning with config 6.

Discussion: The results show that if there are actually one, two, three, or four different locations for the QTL studied, strategy S_2 proposed in this article is able to give a better estimation of the x_i than the use of estimated positions $\hat{x}_i$. The different comparison indicators try to measure the quality of this estimation. They give consistent results. Our method combines different QTL location estimates $\hat{x}_i$, as is usually done in meta-analysis studies even if they manipulate other types of data (*e.g.*, Britten 1996; Allison and Heo 1998; Van Zandt and Mopper 1998; Vøllestad *et al.* 1999). However, these studies deal with what would correspond to only one common QTL location in our case.

TABLE 3

Mean squared error of prediction R^{S_h}, length of the confidence interval at 90% $q(0.90)$ [respectively, 95% $q(0.95)$] computed with 500 simulations in different configurations for both strategies

	Config 1		Config 11		Config 12		Config 6	
Strategies	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
R^{S_h} (*100)	**1.03**	**0.075**	**1.02**	**0.545**	**1.08**	**0.140**	**1.03**	**0.455**
$q(0.90)$	**0.168**	**0.041**	**0.165**	**0.068**	**0.165**	**0.055**	**0.168**	**0.056**
$q(0.95)$	**0.199**	**0.047**	**0.198**	**0.095**	**0.198**	**0.071**	**0.199**	**0.071**
	Config 13		Config 8		Config 14		Config 15	
Strategies	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
R^{S_h} (*100)	*1.01*	*0.885*	**1.02**	**0.520**	1.01	1.010	1.02	1.50
$q(0.90)$	*0.166*	*0.135*	**0.168**	**0.089**	0.167	0.169	0.167	0.203
$q(0.95)$	*0.199*	*0.249*	**0.197**	**0.117**	0.197	0.253	0.196	0.238
	Config 16		Config 17		Config 18		Config 19	
Strategies	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
R^{S_h} (*100)	**0.255**	**0.035**	**0.500**	**0.100**	*2.30*	*1.94*	*4.085*	*3.775*
$q(0.90)$	**0.083**	**0.027**	**0.116**	**0.039**	*0.249*	*0.235*	*0.332*	*0.281*
$q(0.95)$	**0.099**	**0.033**	**0.138**	**0.048**	*0.296*	*0.372*	*0.394*	*0.400*
	Config 20		Config 21		Config 22		Config 23	
Strategies	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
R^{S_h} (*100)	**0.261**	**0.034**	**0.510**	**0.067**	**2.35**	**0.675**	**4.055**	**2.120**
$q(0.90)$	**0.084**	**0.028**	**0.117**	**0.039**	**0.252**	**0.091**	**0.333**	**0.136**
$q(0.95)$	**0.100**	**0.034**	**0.140**	**0.048**	**0.300**	**0.113**	**0.394**	**0.185**
	Config 24		Config 25		Config 26		Config 27	
Strategies	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
R^{S_h} (*100)	**1.26**	**0.078**	**1.295**	**0.845**	**1.295**	**0.262**	**1.35**	**0.980**
$q(0.90)$	**0.184**	**0.038**	**0.186**	**0.083**	**0.186**	**0.059**	**0.188**	**0.109**
$q(0.95)$	**0.225**	**0.047**	**0.231**	**0.161**	**0.231**	**0.083**	**0.232**	**0.178**
	Config 28		Config 29		Config 30		Config 31	
Strategies	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
R^{S_h} (*100)	**1.230**	**0.132**	**1.295**	**0.881**	**1.295**	**0.394**	**1.342**	**1.025**
$q(0.90)$	**0.181**	**0.045**	**0.186**	**0.098**	**0.186**	**0.081**	**0.188**	**0.133**
$q(0.95)$	**0.223**	**0.056**	**0.231**	**0.211**	**0.231**	**0.132**	**0.232**	**0.203**
	Config 32		Config 33		Config 34		Config 35	
Strategies	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
R^{S_h} (*100)	**1.225**	**0.229**	*1.371*	*1.015*	**1.368**	**0.523**	**1.365**	**1.056**
$q(0.90)$	**0.178**	**0.059**	*0.191*	*0.116*	**0.191**	**0.102**	**0.190**	**0.144**
$q(0.95)$	**0.220**	**0.079**	*0.233*	*0.264*	**0.233**	**0.148**	**0.234**	**0.209**

The values are indicated in boldface when S_2 is better than S_1 for all the indicators and in italic when no strategy is better for all the indicators. See Table 2 for a description of configurations.

The theory is developed for independent experiments and known variance. We apply this theory for nonindependent observations in the simulation section and consider the effect of imperfect knowledge of the variance. The quality of the results in these cases shows that the method is robust and that there is no need for a specific theory to take nonindependence and estimation of the variances into account.

A particular situation is the case where two different QTL are detected on the same chromosome for the same trait and in the same experiment. In this case, considering the two QTL as independent will not take the previous information into account.

Imagine a situation where we have all the markers and phenotypic information for the different experiments and a join map of all the markers. It would then

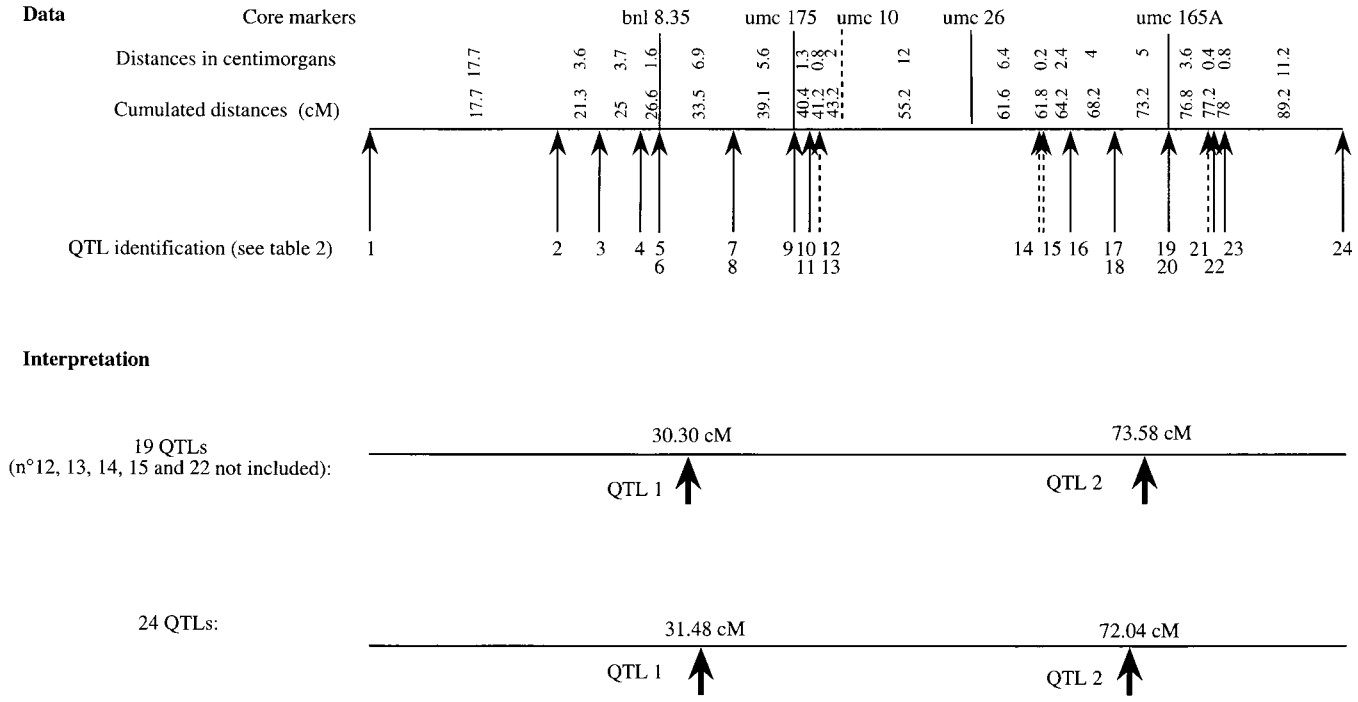

Figure 1.—QTL related to yield on linkage group 3 of the maize genome.

be possible to perform a global linkage analysis and to look for common QTL in each position as in Haley *et al.* (1994) or Rebai and Goffinet (1993). The question of distinguishing between one, two, three or more QTL becomes a different problem in this case. As shown, for example, in Goffinet and Mangin (1998), the distinction between these different hypotheses is not easy.

The expressions for AIC*(k) are given for $k = 1, 2, 3, 4, n$. It would be interesting to obtain this expression for $k > 4$, but as noted previously, it would only be useful for values of $n > 40$ and for a chromosome whose length is >2 M. However, according to the dense linkage maps existing nowadays for many different species, mean chromosome lengths never exceed this value.

AN APPLICATION USING THE MAIZE GENOME DATABASE

Using the maize database (at http://www.agron.missouri.edu), we collected the data concerning QTL related to yield and located on linkage group 3. We looked through the original publications and were able to construct a "consensus" map, where all the QTL could be localized. This map was based on core markers that were present in the different publications. The distances between two markers could differ between publications but were quite similar: we took the average values for our map. A total of 24 QTL could be detected; their position is given in Figure 1, and their description is in Table 4. Five of them were mapped relative to marker umc10 (QTL numbers 12, 13, 14, 15, and 21), whose localization on the map was not precise. At this stage we discard them from the analysis and use only the other 19 QTL.

Our linkage data span from bin 3.4 to bin 3.6 according to the nomenclature of the maize database.

Looking through the different studies in which the data were collected, we were able to estimate confidence intervals for the majority of QTL positions. If we consider these positions to be normally distributed and a confidence interval $C(90)$ of 90%, the standard deviations $\sigma_{E,i}$ of the different QTL can be estimated as $C(90) = 2 \times 1.645 \times \sigma_{E,i}$ cM. These values are given in Table 4. For those QTL where no confidence interval could be evaluated, the value of $\sigma_{E,i}$ was taken as 6, 10, or 15 cM, corresponding to confidence intervals of 20, 33, or 50 cM, the second value equaling the mean of our estimated confidence intervals. The QTL number 1 is quite far from the others. This QTL must have a large variance: its position is likely to be inaccurately estimated, since it is located in an interval of 42.6 cM without any marker and 16 cM apart from the nearest marker (Veldboom and Lee 1996). For these reasons, we attributed a $\sigma_{E,i}$ of 20 cM to this QTL for further analysis.

We first tested our model with 19 QTL, with a σ_E of 10 cM; then we included the 5 QTL localized relative to marker umc10 in the data, that is, 24 QTL with the different values of $\sigma_{E,i}$. The results are given in Table 5 and are discussed in the next section.

Discussion: In Table 5, the underlined number is the best value of the criterion. In all cases, the model with two positions is favored by the criteria, whatever the value of σ_E and the number of QTL considered.

TABLE 4

QTL related to yield on linkage group 3 of the maize genome

Identification in Figure 1	Trait	Estimated $\sigma_{E,i}$ (cM)	Reference
1	Kernel weight	[a]	Veldboom and Lee (1996)
2	Cob diameter	16	Beavis *et al.* (1994)
3	Plant height	8	Beavis *et al.* (1994)
4	Test weight	[a]	Ajmone-Marsan *et al.* (1995)
5	Plant height	9.6	Beavis *et al.* (1994)
6	Kernel row number	6.5	Austin and Lee (1996)
7	Plant height	12	Beavis *et al.* (1991)
8	Plant height	12	Beavis *et al.* (1991)
9	Plant height	6.7	Schön *et al.* (1993)
10	Kernel weight	9.8	Austin and Lee (1996)
11	Kernel weight	15.3	Veldboom and Lee (1994)
12	Grain weight	[a]	Maize Database, CIMMYT (1994)[b]
13	Plant height	[a]	Maize Database, CIMMYT (1994)[b]
14	Test weight	8	Beavis *et al.* (1994)
15	Plant height	[a]	Maize Database, CIMMYT (1994)[b]
16	Ear diameter	5.1	Veldboom and Lee (1994)
17	Ear number per plant	8.5	Veldboom and Lee (1994)
18	Ear number per plant	[a]	Veldboom and Lee (1996)
19	Ear diameter	7.6	Austin and Lee (1996)
20	Kernel weight	7.6	Austin and Lee (1996)
21	Plant height	[a]	Maize Database, CIMMYT (1994)[b]
22	Ear number per plant	5.5	Austin and Lee (1996)
23	Plant height	14.9	Schön *et al.* (1994)
24	Ear length	6.8	Veldboom and Lee (1994)

[a] No information available in the reference.
[b] http://www.agron.missouri.edu:80/cgi_bin/sybgw_mdb/mdb3/reference/67081.

The 19 QTL are well represented by two real QTL located on positions 30.30 (QTL 1) and 73.58 (QTL 2; see large arrows in Figure 1). QTL 1–11 would be representative of a first QTL at 30.30 cM, QTL 16–24 would be a second one at 73.58 cM (Figure 1). The trait affected by QTL 1 is mainly plant height whereas QTL 2 mainly affects ear traits (Table 4).

When the QTL located relative to locus umc10 are included in the analysis, the results are not much affected [Table 5 (24 QTL) and Figure 1]. The estimation of the positions for the models with two QTL are close to those estimated with 19 QTL.

CONCLUSION

As the number of studies concerning QTL detection increases, articles dealing with the use of results from several studies concentrating on different species (Kearsey and Farquhar 1998) or on a single one (Khavkin and Coe 1997, 1998) are now available. A major step toward a more accurate identification of a QTL consists of finding the proper candidate gene.

A candidate gene for a given trait is a sequence of a gene of a known biological function involved with the development or physiology of the trait. However, the likelihood that a given candidate gene corresponds to a given QTL is very small; there are many possible genes, and the candidate can be chosen in two different ways. First, the candidate gene can be chosen on an *a priori* belief that, due to its function, the gene is associated with the trait of interest. Second, the gene can be suspected to be the candidate because it is located in the area of the QTL: this is a *positional comparative candidate-gene analysis* (Rotschild and Soller 1997). Unless the QTL position confidence interval is very narrow, the initial candidate gene can be incorrect. To increase the power of detection, the confidence interval must be narrowed, or the results from several genome-wide surveys must be combined (Keightley *et al.* 1998), which is precisely what we suggest in this study. Gathering QTL data together should be a good way to obtain a better estimation of a QTL position and thus to specify a colocation with a candidate gene. Moreover, the reduction of the confidence interval associated with QTL location is an important goal (see, for instance, Kearsey and Farquhar 1998), so the reduction provided by the method presented in this article is therefore of advantage.

In the maize genome, functional clusters were found associating QTL and genes for growth, development, and stress response. The genomic location of the QTL used in our example (chromosome 3, bins 4–6) contains, for instance, genes for auxin and ABA sensors,

TABLE 5

Application of the models to the maize data

19 QTL (QTL nos. 12, 13, 14, 15, and 22 not included)					
Unknown $\sigma_{E,i}$ = 10 cM					
Models	$k = 19$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC*(k)	−15.41	60.33	−45.94[a]	−37.26	−32.35
Estimated positions (cM) (QTL no. 1 being at 0 cM)					
$k = 1$	54.86				
$k = 2$	30.30	73.58			
$k = 3$	23.59	38.19	73.58		
$k = 4$	23.59	38.19	68.40	81.34	
24 QTL					
Unknown $\sigma_{E,i}$ = 6 cM					
Models	$k = 24$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC*(k)	−23.39	72.97	−55.96	−48.45	−45.09
Unknown $\sigma_{E,i}$ = 10 cM					
Models	$k = 24$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC*(k)	−17.26	44.10	−56.08	−45.91	−39.53
Unknown $\sigma_{E,i}$ = 15 cM					
Models	$k = 24$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC*(k)	−12.40	37.94	−53.34	−42.50	−34.87
Estimated positions (cM)					
$k = 1$	55.09				
$k = 2$	31.48	72.04			
$k = 3$	31.48	64.59	78.36		
$k = 4$	23.59	38.87	64.59	78.3	

[a] Underlined, best value of the criterion.

genes for reduced or distorted growth of shoot, leaf, male and female inflorescence, loci for reduced plant vigor, and loci for a transcription binding factor (Khavkin and Coe 1997). Moreover, according to these authors, chromosomes 1 and 3 seem to carry 40% of all developmental genes. Thanks to the associated skills of physiologists and geneticists, as a growing number of genes are mapped and as their function is increasingly well elucidated, the number of potential candidate genes is increasing. Tools are needed to determine the relevant candidate, and specifying a colocation is one of them.

The correspondence of QTL across genomes of different species is illustrated by several studies, for instance between different Brassica species and Arabidopsis, where the potential of integrating QTL analysis with comparative studies and candidate loci suggested by the synteny Brassica/Arabidopsis is highlighted for flowering traits (Bohuon *et al.* 1998). Comparative mapping yields a more comprehensive list of QTL than can be obtained from any individual population (Lin *et al.* 1995), so a meta-analysis approach can be performed within a species, and the result obtained can be tested on another species. The existence of model species on which studies are concentrated facilitates meta-analysis. The conservation of gene order between species provides a framework for the comparative analysis of complex phenotypes (Paterson *et al.* 1995). Using meta-analysis to extract meaningful results for a particular species may in this way have a greater impact.

LITERATURE CITED

Ajmone-Marsan, P., G. Monfredini, W. F. Ludwig, A. E. Melchinger, P. Franceschini *et al.*, 1995 In an elite cross of maize a major quantitative trait locus controls one-fourth of the genetic variation for grain yield. Theor. Appl. Genet. **90:** 415–424.

Allison, D. B., and M. Heo, 1998 Meta-analysis of linkage data under worst-case conditions: a demonstration using the human OB region. Genetics **148:** 859–865.

Austin, D. F., and M. Lee, 1996 Comparative mapping in F2:3 and F6:7 generations of quantitative trait loci for grain yield and yield components in maize. Theor. Appl. Genet. **92:** 817–826.

Beavis, W. D., D. Grant, M. Albertsen and R. Fincher, 1991 Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. Theor. Appl. Genet. **83:** 141–145.

Beavis, W. D., O. S. Smith, D. Grant and R. Fincher, 1994 Identification of quantitative trait loci using a small sample of topcrossed and F4 progeny from maize. Crop Sci. **34:** 882–896.

Bohuon, E. J. R., L. D. Ramsay, J. A. Craft, A. E. Arthur, D. F. Marshall *et al.*, 1998 The association of flowering time quantitative trait loci with duplicated regions and candidate loci in *Brassica oleracea*. Genetics **150:** 393–401.

Britten, H. B., 1996 Meta-analysis of the association between multilocus heterozygosity and fitness. Evolution **50:** 2158–2164.

Byrne, P. F., M. B. Berlyn, E. H. Coe, G. L. Davis, M. L. Polacco *et al.*, 1995 Reporting and accessing QTL information in USDA's

Maize Genome Database. J. Quant. Trait Loci (at http://probe.nalusda.gov:8000/otherdocs/jqtl/index.html).

Fulton, T. M., T. Beckbunn, D. Emmatty, Y. Eshed, J. Lopez *et al.*, 1997 QTL analysis of an advanced backcross of *Lycopersicon peruvianum* to the cultivated tomato and comparisons with QTLs found in other wild species. Theor. Appl. Genet. **95:** 881–894.

Goffinet, B., and B. Mangin, 1998 Comparing methods to detect more than one QTL on a chromosome. Theor. Appl. Genet. **96:** 628–633.

Haley, C. S., S. A. Knott and J. M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136:** 1195–1207.

Hedges, L. V., and I. Olkin, 1985 *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL.

Hyne, V., and M. J. Kearsey, 1995 QTL analysis: further uses of 'marker regression'. Theor. Appl. Genet. **91:** 471–476.

Jansen, R. C., 1996 Complex plant traits: time for polygenic analysis. Trends Plant Sci. **1:** 89–94.

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

Keightley, P. D., K. H. Morris, A. Ishikawa, V. M. Falconer and F. Oliver, 1998 Test of candidate gene-quantitative trait locus association applied to fatness in mice. Heredity **81:** 630–637.

Khavkin, E., and E. H. Coe, 1997 Mapped genomic locations for developmental functions and QTLs reflect concerted groups in maize (*Zea mays* L.). Theor. Appl. Genet. **95:** 343–352.

Khavkin, E., and E. H. Coe, 1998 The major quantitative trait loci for plant stature, development and yield are general manifestations of developmental gene clusters. Maize Newslett. **72:** 60–66.

Kowalski, S. P., T. H. Lan, K. A. Feldmann and A. H. Paterson, 1994 Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization. Genetics **138:** 499–510.

Lander, E., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting results. Nat. Genet. **11:** 241–247.

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

Mangin, B., B. Goffinet and A. Rebai, 1994 Constructing confidence intervals for QTL location. Genetics **138:** 1301–1308.

Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln *et al.*, 1988 Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335:** 721–726.

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.

Paterson, A. H., T. H. Lan, K. P. Reischmann, C. Chang, Y. R. Lin *et al.*, 1996 Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nat. Genet. **14:** 380–382.

Rebai, A., and B. Goffinet, 1993 Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor. Appl. Genet. **86:** 1014–1022.

Rotschild, M. F., and M. Soller, 1997 Candidate gene analysis to detect genes controlling traits of economic importance in domestic livestock. Probe **8:** 13–20.

Sakamoto, Y., M. Ishiguro and G. Kitagawa, 1986 *Akaike Information Criterion Statistics*. KTK Scientific Publishers, Tokyo.

Schön, C. C., M. Lee, A. E. Melchinger, W. D. Guthrie and W. L. Woodman, 1993 Mapping and characterization of quantitative trait loci affecting resistance against second-generation European corn borer in maize with the aid of RFLPs. Heredity **70:** 648–659.

Schön, C. C., A. E. Melchinger, J. Boppenmaier, E. Brunklaus-Jung *et al.*, 1994 RFLP mapping in maize: quantitative trait loci affecting testcross performances of elite european flint lines. Crop Sci. **34:** 378–389.

Titterington, D. M., A. F. M. Smith and U. E. Makov, 1985 *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

Van Zandt, P., and S. Mopper, 1998 A meta-analysis of adaptive deme formation in phytophagous insect populations. Am. Nat. **152:** 595–604.

Veldboom, L. R., and M. Lee, 1994 Molecular-marker-facilitated studies of morphological traits in maize. II. Determination of QTLs for grain yield and yield components. Theor. Appl. Genet. **89:** 451–458.

Veldboom, L. R., and M. Lee, 1996 Genetic mapping of quantitative trait loci in maize in stress and nonstress environments. I. Grain yield and yield components. Crop Sci. **36:** 1310–1319.

Vøllestad, L. A., K. Hindar and A. P. Møller, 1999 A meta-analysis of fluctuating asymmetry in relation to heterozygosity. Heredity **83:** 206–218.

Communicating editor: C. Haley