Quantitative Trait Loci Mapping in F_2 Crosses Between Outbred Lines

Miguel Pérez-Enciso and Luis Varona

Centre UdL-IRTA, Area de Producció Animal, 25198 Lleida, Spain

Manuscript received March 16, 1999
Accepted for publication January 10, 2000

ABSTRACT

We develop a mixed-model approach for QTL analysis in crosses between outbred lines that allows for QTL segregation within lines as well as for differences in mean QTL effects between lines. We also propose a method called "segment mapping" that is based in partitioning the genome in a series of segments. The expected change in mean according to percentage of breed origin, together with the genetic variance associated with each segment, is estimated using maximum likelihood. The method also allows the estimation of differences in additive variances between the parental lines. Completely fixed random and mixed models together with segment mapping are compared via simulation. The segment mapping and mixed-model behaviors are similar to those of classical methods, either the fixed or random models, under simple genetic models (a single QTL with alternative alleles fixed in each line), whereas they provide less biased estimates and have higher power than fixed or random models in more complex situations, *i.e.*, when the QTL are segregating within the parental lines. The segment mapping approach is particularly useful to determining which chromosome regions are likely to contain QTL when these are linked.

QUANTITATIVE traits arise from the joint action of the environment and multiple genes, usually called quantitative trait loci (QTL). The wide availability of DNA markers scattered along the genome, together with recently developed statistical methods, has spurred the massive search for QTL in any species of interest. Crosses between highly divergent lines are a powerful experimental design for this purpose (Lynch and Walsh 1998). The optimum situation in a F_2 design occurs when all genes affecting the trait of interest are diallelic with the alternative alleles fixed in each parental line. Although in annual plant species and some lab animals highly inbred lines that may fulfill this condition have been developed, outbred parental populations are normally the only genetic material available in domestic animals (*e.g.*, Andersson *et al.* 1994) or trees (*e.g.*, Grattapaglia *et al.* 1995), as well as in allogamous wild species (*e.g.*, Hunt *et al.* 1998). The QTL analysis of crosses between outbred populations poses two main statistical problems (reviews in Bovenhuis *et al.* 1997; Hoeschele *et al.* 1997; Elsen *et al.* 1999). The first one concerns the validity of the genetic model assumed in the analysis. The second one is related to accounting for the variation in the rest of the genome when fitting a QTL model at a particular position.

The usual model for analyzing F_2 crosses (Lander and Botstein 1989; Haley and Knott 1992) is based on estimating the QTL effect from the phenotypic differences between individuals according to the estimated percentage of breed origin at a given position, assuming that alternative alleles are fixed in each parental line. We call this model the fixed model. Yet, the fact that heritability for a given trait is nonzero, as in most outbred lines, implies that there exists additive variation within lines and thus not all alleles affecting the trait can be fixed. There are also methods that allow for QTL segregation where the QTL effect is modeled as a normally distributed random variable with mean zero and variance to be estimated. This is the random model. The random model strategy has been put forward by several authors in the context of the analysis of outbred populations (Fernando and Grossman 1989; Goldgar 1990; Xu and Atchley 1995; Grignola *et al.* 1996). The QTL variance is estimated by assessing the degree of phenotypic similarity between relatives according to the probability of sharing identical by descent alleles at specified positions. But the random model does not seem appropriate for the analysis of F_2 crosses because no particular distinction is made between allele breed origin in current implementations. A strategy similar to the random model is the within-family analyses, where each family (*e.g.*, descendants of each sire) is analyzed separately and the results pooled (*e.g.*, Knott *et al.* 1996). However, this approach will tend to have small power when the family size and the QTL effect decrease.

A mixed-model approach that accounts for variation both between and within lines is thus the most appropriate strategy for analyzing F_2 crosses between outbred lines. Goddard (1992) proposed a QTL mixed-model strategy for genetic evaluation that can potentially be applied to crosses between outbred lines, but marker information is used only to model covariances between

Corresponding author: Miguel Pérez-Enciso, Station d'Amélioration Génétique des Animaux, INRA, BP 27, 31326 Castanet-Tolosan Cedex, France. E-mail: mperez@toulouse.inra.fr

QTL effects, not means, and the method does not account for differences in means and heritabilities between breeds in the genetic covariance matrix of crossed individuals; it is also assumed that marker phases are known in constructing the relationship matrix. Lo *et al.* (1993) developed the covariance between relatives in crosses between outbred populations for a number of unlinked loci and without marker information, whereas Wang *et al.* (1998) studied the case of a single marker and a QTL in a genetic evaluation context.

The problem of accounting for the genetic variation in the rest of the genome has been addressed by proposing the use of cofactors ("composite interval mapping"; Jansen 1993; Zeng 1993), but it would be desirable to have a methodology that addresses the issue more generally. Other authors have included a polygenic effect in addition to the fixed QTL effect (*e.g.*, Fernando and Grossman 1989), but this does not allow for the fact that not all the genome contributes equally to the genetic variation and implies that this polygenic component is unlinked to the QTL of interest.

In this work we derive the genetic covariance matrix in crosses between outbred lines allowing for any number of linked markers and QTL, thus permitting a general QTL analysis of F_2 crosses. This mixed model allows for more flexible genetic models than current strategies. We also propose a method, "segment mapping," aimed at accounting for the variation in the whole genome simultaneously. The method also allows us to test genetic variance differences between breeds. A simulation study is carried out to compare the performance of segment mapping and mixed model mapping with classical methods, *i.e.*, a genome scan using fixed or random models.

THEORY

The breeding value of an individual is, by definition, twice the average performance of an infinite number of its offspring when mated to a random sample of spouses from the same population. The starting point is the assumption that the breeding values (g) of two outbred populations A and B are normally distributed $g_A \sim N(\mu + \Delta/2, \sigma^2_A)$ and $g_B \sim N(\mu - \Delta/2, \sigma^2_B)$, respectively. The phenotypic difference between breeds for the trait of interest is thus Δ. Genetic variation within breeds is assumed to be caused by an indeterminate number of loci in genetic equilibrium with additive action. Further, consider that the whole genome is divided in n_{seg} segments and that a vector containing the additive genetic values from the population of breed A can be expressed as $\mathbf{g}_A = \Sigma_{s=1}^{n_{seg}} \mathbf{g}_{A,s}$, where $\mathbf{g}_{A,s}$ is the contribution of segment s to total breeding value, and $\text{Var}(\mathbf{g}_A) = \Sigma_{s=1}^{n_{seg}} \text{Var}(\mathbf{g}_{A,s}) = \Sigma_{s=1}^{n_{seg}} \mathbf{G}_{A,s}$ because of linkage equilibrium. In the absence of molecular information, $\text{Var}(\mathbf{g}_A)$ is the well-known additive relationship matrix and $\mathbf{G}_{A,s}$ is the same for all segments (weighed by the segment's length). However, the availability of marker information makes it possible to compute the probabilities of identity by descent at particular positions of interest (*e.g.*, Fernando and Grossman 1989).

The goal of the approach presented here is to estimate, conditional on marker information, the contribution of each segment to total genetic variance/covariance between the F_2 individuals and to ascertain the expected phenotypic mean of individuals according to the percentage of breed origin in each particular segment. A reasonable strategy would be to include loci of similar effect in the same segment but the theory developed is valid for any partition strategy.

Assume that trait performance has been recorded in a F_2 cross population derived from breed A and B and that parental, F_1, and F_2 individuals have been genotyped for a series of markers. A general explanatory model of the F_2 records is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg}_{F2} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is a $N \times 1$ vector containing the F_2 phenotypes, $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating observations to the vector of fixed effects ($\mathbf{b}$) and additive genetic values ($\mathbf{g}$), respectively, and $\mathbf{e}$ contains the residuals. In the following we refer only to breeding values in the F_2 population and thus the subscript is omitted for brevity. The distribution of the random variables in (1) is

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{g} \\ \mathbf{e} \end{pmatrix} = \left[\begin{pmatrix} \mathbf{Xb} + \mathbf{Q\Delta} \\ \mathbf{Q\Delta} \\ \mathbf{O} \end{pmatrix}, \begin{pmatrix} \mathbf{V} & \mathbf{GZ}' & \mathbf{R} \\ \mathbf{ZG} & \mathbf{G} & \mathbf{O} \\ \mathbf{R} & \mathbf{O} & \mathbf{R} \end{pmatrix} \right], \tag{2}$$

where $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, $\mathbf{G}$ is the genetic covariance matrix conditional on marker information as specified below, $\mathbf{R} = \mathbf{I}\sigma^2_e$, $\mathbf{I}$ being a diagonal unit matrix and σ^2_e the residual variance, $\mathbf{Q}$ is a $N \times n_{seg}$ matrix with elements $q_{i,s} = (\Sigma_{h=1}^2 p^h_{i,s} - 1)/2$, $p^h_{i,s}$ is the average probability of segment s from individual i and haplotype h being of breed origin A, and $\Delta = \{\Delta_s, s = 1, n_{seg}\}$, *i.e.*, a vector containing the average differences between individuals carrying an A breed origin segment s minus those carrying a B origin segment. Further, $\text{Var}(\mathbf{g}) = \mathbf{G} = \Sigma_{s=1}^{n_{seg}} \mathbf{G}_s$, assuming linkage equilibrium in the parental populations and that markers are informative (see the appendix). Otherwise, the $\mathbf{g}_s$ from different within-chromosome segments will be correlated. The matrix $\mathbf{G}_s$ contains elements $\text{Var}(g_{i,s})$ in the diagonal and $\text{Cov}(g_{i,s}, g_{i',s})$ in the off-diagonal. It is shown in the appendix that the variance of breeding values of F_2 individuals, conditional on marker information, is approximately

$$\text{Var}(g_i) \approx \sum_{s=1}^{n_{seg}} \text{Var}(g_{i,s}) \approx \sum_{s=1}^{n_{seg}} \sum_{h=1}^{2} [p^h_{i,s}\, \sigma^2_{A,s} + (1 - p^h_{i,s})\sigma^2_{B,s}], \tag{3}$$

where $\sigma^2_{A,s}$ and $\sigma^2_{B,s}$ are the genetic variances contributed by segment s within parental populations A and B, re-

spectively. Thus, the genetic variance of F_2 individuals, conditional on marker information, is a weighted average of the genetic variances in the pure breeds. It is important to realize that the segregation variance (Wright 1968) can be neglected in (3) because the expression above is the genetic variance conditional on marker information. Equation 3 would be exact if the breed origin along the whole genome could be identified without error. Suppose that a subset of individuals with its whole genome of origin A could be identified in an infinitely large F_2 population; the genetic variance of these individuals would be σ^2_A, exactly that of the founder breed A. The additive genetic covariance between F_2 individuals is

$$\mathrm{Cov}(g_i, g_{i'}) = \sum_{s=1}^{n_{\mathrm{seg}}} \sum_{h=1}^{2} [\rho^h_{A(i,i'),s}\, \sigma^2_{A,s} + \rho^h_{B(i,i'),s}\, \sigma^2_{B,s}], \quad (4)$$

where $\rho^h_{A(i,i'),s}$ ($\rho^h_{B(i,i'),s}$) is the probability of individuals i and i' having identical by descent alleles of breed origin A (B) at segment s and haplotype h. Equation 4 shows that two individuals can share alleles identical by descent of breed origin A or B and that the total genetic covariance is a weighted average of both probabilities.

The model in (1) and (2) together with (3) and (4) provides the general framework to analyze F_2 populations using standard mixed-model theory and molecular markers. These equations account for the fact that the average effect of alleles can be different between breeds, but also that there can simultaneously exist a QTL segregation within breeds. The average difference in allelic effects between both breeds is included as a fixed effect through $\mathbf{Q}\Delta$, whereas the additional variation within breeds is allowed through $\mathbf{G}$. The usual genome scan/regression strategy means that model (1) is fitted with an infinitesimally small segment (= 1 QTL) in successive positions assuming $\sigma^2_A = \sigma^2_B = 0$. If only one QTL is fitted at a time, the matrix $\mathbf{Q}$ is a vector with coefficients as in, *e.g.*, Haley and Knott (1992). In contrast, σ^2_A and σ^2_B are larger than zero for those QTL with alleles not fixed in the parental populations. The simple fixed model is not appropriate because not all differences between individuals due to that segment are fully accounted for by Δ_s. Note that it is straightforward to accommodate that alleles are fixed in only one of the two breeds.

Molecular information is used to calculate $p^h_{i,s}$, $\rho^h_{A(i,i'),s}$, and $\rho^h_{B(i,i'),s}$. Note that only the breed origin probabilities are involved in obtaining $p^h_{i,s}$, whereas the identity by descent probabilities between marker alleles are required to compute $\rho^h_{A(i,i'),s}$ and $\rho^h_{B(i,i'),s}$. If two F_2 individuals do not have any common ancestor, $\rho^h_{A(i,i'),s} = \rho^h_{B(i,i'),s} = 0$ necessarily for all segments. But if both are homozygous for marker alleles that can be traced back unambiguously to breed A, $p^h_{i,s} = p^h_{i',s} = 1$, for that particular position, and could differ for other segments. In an ideal situation of infinite number of informative markers, these quantities are easy to compute. For instance the fraction of the genome of origin A is

$$p_i = \frac{1}{2L} \sum_{s=1}^{n_{\mathrm{seg}}} L_s \sum_{h=1}^{2} p^h_{i,s} = \frac{1}{2L} \sum_{h=1}^{2} \int_0^L \delta^h_i(x)\partial(x),$$

where $\delta^h_i(x)$ is a Dirac function taking value 1 if haplotype h at point x is of origin A and zero otherwise, and L_s and L are the segment length and the total length of the genome in morgans, respectively. If markers are not completely informative or the map is not infinitely dense, several options can be employed. Note that only the breed origin probabilities are needed to compute $p^h_{i,s}$ and, *e.g.*, the method in Haley *et al.* (1994) can be employed. In contrast, the identity by descent probabilities need to be obtained to compute $\rho^h_{A(i,i'),s}$ and $\rho^h_{B(i,i'),s}$. These are given by Fernando and Grossman (1989) for a QTL linked to a single marker. Grignola *et al.* (1996) provide a more general algorithm. Monte Carlo Markov chain methods like that in Heath (1997) can also be employed. We have developed a Monte Carlo Markov chain algorithm because of its flexibility and because it takes into account all available information, considering simultaneously the molecular information from all individuals. The procedure is based on a Gibbs sampler that samples and updates successively the phase of markers for every individual conditional on the phase of its spouse, parents, and offspring. For each Gibbs iteration, crossover locations for an individual's genome are simulated conditional on its current phase and the phase of its parents. A noninterference Haldane's mapping function is used. Once all crossover locations are simulated, the parentage between all individuals is obtained by tracing back the genome origins at the specified segments. The total relationship is obtained by averaging the relationship over Gibbs iterates.

Parameter estimates of $\mathbf{b}$, Δ_s, $\sigma^2_{A,s}$ and $\sigma^2_{B,s}$ can obtained by maximum likelihood using the Simplex algorithm. This algorithm is a derivative-free method and requires only the logarithm of the likelihood, *i.e.*,

$$L = -\tfrac{1}{2}[\mathrm{Constant} + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{Xb} - \mathbf{Q}\,\Delta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb} - \mathbf{Q}\Delta)].$$

It should be noted that the average $\mathbf{G}$ over Gibbs iterates is used here and that the method can be, potentially, improved by marginalizing with respect to $\mathbf{G}$, $\mathbf{b}$, Δ, and σ^2, as in a Bayesian framework.

SIMULATION

We carried out a simulation study to test the performance of segment mapping and mixed-model scan *vs.* standard strategies. The F_2 pedigree consisted of 5 parental sires from breed A, each mated to 2 dams of breed B that produced 5 F_1 sires (1 per parental sire) and 40 F_1 dams (4 per parental dam). The number of F_2 offspring was 400. A 60-cM chromosome was simu-

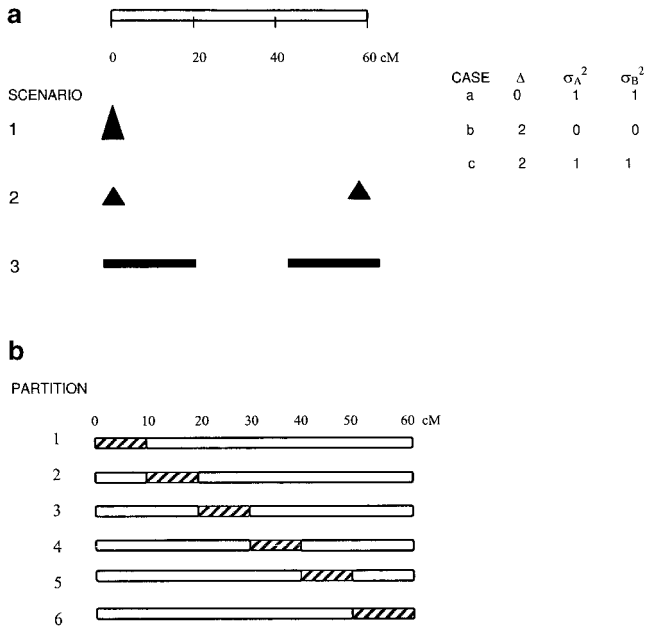

Figure 1.—(a) Scheme of the genetic scenarios and cases considered. The open bar represents the chromosome with numbers at the marker positions. The solid arrows/bars indicate the positions of the QTL for each scenario; the thickness is proportional to the effect of the QTL. The cases considered within each scenario are shown within the frame, where Δ is the phenotypic difference between breeds A and B, and σ^2_A and σ^2_B are the genetic variance in breeds A and B, respectively. Only case c was considered in scenarios 2 and 3. (b) Scheme of chromosome partitions used with segment mapping in scenarios 1 and 2. The two segments considered are hatched and open, respectively.

lated, and completely informative markers (*i.e.*, each line had different marker alleles and as many alleles as founder individuals were generated) were located at positions 0, 20, 40, and 60 cM. Three genetic scenarios as depicted in Figure 1 were considered. A single telomeric locus explained all genetic differences between lines in scenario 1, and there were two telomeric loci at positions 0 and 60 cM in scenario 2. In scenario 3 there were two spaced clusters of 20 genes each, and the loci were of equal effect located every centimorgan in positions 1–20 and 41–60 cM. Three distinct cases were studied in scenario 1. First (case a) $\sigma^2_A = \sigma^2_B = \sigma^2_e = 1$ and $\Delta = 0$; *i.e.*, this is equivalent to an outbred population, as there are no expected phenotypic differences according to allele origin. In case b the alleles were fixed within breed ($\sigma^2_A = \sigma^2_B = 0$), $\sigma^2_e = 2$, and $\Delta = 2$. This is the current genetic model assumed in analyzing F_2 crosses. And finally (case c), $\sigma^2_A = \sigma^2_B = \sigma^2_e = 1$ and $\Delta = 2$; *i.e.*, there are phenotypic differences between breeds but still there exists additive variance within the parental populations. This is the situation occurring in F_2 crosses between divergent outbred populations. It was the only case considered for genetic scenarios 2 and 3. Thirty replicates per model and case were run. The allele effects of the founder individuals were simulated according to its expected distribution, *e.g.*, for breed A, $N[(\mu + \Delta/2)/(2\, n_{loci}), \sigma^2_A/(2\, n_{loci})]$, where n_{loci} is the number of loci, *i.e.*, 1, 2, and 40 for scenarios 1, 2, and 3, respectively. Phenotypes were generated by summing the allele effects of the F_2 individual and adding a residual normal variate of mean zero and variance 1 (case a and c) or 2 (case b). There was no sexual dimorphism and the general mean was the only fixed effect considered.

Four methods of analysis were compared:

Segment mapping: The chromosome was divided into two segments, a 10-cM segment (genetic scenarios 1 and 2) or 20 cM (scenario 3) and a segment comprising the rest of the chromosome. The model was

$$\begin{aligned}\mathbf{y} &= \mu + \mathbf{g}_s + \mathbf{g}_{\bar{s}} + \mathbf{e}\\ &= \mu + \mathbf{p}_s\,\Delta_s + \mathbf{u}_s + \mathbf{p}_{\bar{s}}\,\Delta_{\bar{s}} + \mathbf{u}_{\bar{s}} + \mathbf{e}, \qquad (5)\end{aligned}$$

where the subscript $\bar{s}$ is used to indicate the complement of segment s (here the rest of the chromosome). It was assumed that genetic variances were equal in both breeds and a single variance component was fitted per segment ($\sigma^2_{A,s} = \sigma^2_{B,s} = \sigma^2_s$ and $\sigma^2_{A,\bar{s}} = \sigma^2_{B,\bar{s}} = \sigma^2_{\bar{s}}$). Above, $\mathbf{g}_s$ is split for convenience into its mean ($\mathbf{p}_s\,\Delta_s$), where $\mathbf{p}_s$ is a vector with elements $(p^1_{i,s} + p^2_{i,s})/2$, and a random genetic variable ($\mathbf{u}_s$) with mean zero. Thus,

$$\begin{pmatrix}\mathbf{u}_s\\ \mathbf{u}_{\bar{s}}\\ \mathbf{e}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{G}_s & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_{\bar{s}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{pmatrix}\right].$$

Several segment partitions were considered. In genetic scenarios 1 and 2, the 10-cM segment was shifted along the chromosome and a total of six analyses were considered, *i.e.*, the first partition consisted of segments at positions 1–10 cM and 11–60 cM; second partition, segments 11–20 cM and the rest (1–10, 21–60 cM); and so on. A scheme of the partitions is in Figure 1b. A similar strategy was followed for genetic scenario 3, except that three partitions of 20 and 40 cM were considered; *i.e.*, the first partition comprised segments 1–20 and 21–60 cM. Note that it is not necessary to establish these successive partitions but it facilitates the comparison with genome scan strategies.

Mixed model: The point model was

$$\mathbf{y} = \mu + \mathbf{g}_s + \mathbf{e} = \mu + \mathbf{p}_s\,\Delta_s + \mathbf{u}_s + \mathbf{e}, \qquad (6)$$

where $\mathbf{u}_s \sim N(0, \mathbf{G}_s)$ as above. This model was fitted in 10-cM intervals for genetic scenarios 1 and 2 and in intervals of 20 cM for scenario 3. The relationship matrix $\mathbf{G}_s$ contains the average relationships in that particular interval. The probabilities $\mathbf{p}_s$ used were the average probabilities in the intervals considered. Note that the common strategy is to compute point probabilities, *e.g.*, every centimorgan, but this has a negligible effect on the results given the small size of the interval and allows us

to compare segment mapping with the mixed model and the two other strategies below.

Random model: The point model was

$$\mathbf{y} = \mu + \mathbf{u}_s + \mathbf{e}, \quad (7)$$

Fixed model: The point model was

$$\mathbf{y} = \mu + \mathbf{p}_s \Delta_s + \mathbf{e}. \quad (8)$$

Random and fixed models were fitted in identical intervals as in the mixed-model strategy.

The relationship matrices and $\mathbf{p}_s$ were obtained after 1000 iterates of the Gibbs sampling scheme. The parameters were estimated in all cases by maximum likelihood using a Simplex algorithm. At each genome partition (segment mapping) or interval position (mixed, random, and fixed models), the likelihood ratio (LR_0) comparing models (5), (6), (7), or (8) *vs.* $\mathbf{y} = \mu + \mathbf{e}$ was computed. In addition, the segment mapping model (5) was compared *vs.* model

$$\mathbf{y} = \mu + \mathbf{g}_{\bar{s}} + \mathbf{e}$$

for each segment partition (LR_s); *i.e.*, the null hypothesis (H_0) tested is that there is no genetic effect in the 10-cM segment (hatched segments in Figure 1b). The likelihood ratios are asymptotically distributed as a chi square with degrees of freedom the difference in number of parameters between models tested. Degrees of freedom are then 1 for $LR_{0,RM}$ (the H_0 in the random model is that σ^2_s is 0) and $LR_{0,FM}$ (the H_0 in the fixed model is that Δ_s is 0), 2 in $LR_{0,MM}$ (the H_0 in the mixed model is that both Δ_s are 0), 4 in $LR_{0,SM}$ (the H_0 for segment mapping is that all Δ_s, $\Delta_{\bar{s}}$, σ^2_s, and $\sigma^2_{\bar{s}}$ are 0), and 2 for LR_s in segment mapping (the H_0 is that Δ_s and σ^2_s are 0). To study the empirical null distribution of the different LR, we simulated 200 replicates under the null hypothesis ($\Delta = \sigma^2_A = \sigma^2_B = 0$).

RESULTS

Table 1 shows the statistics corresponding to the empirical (simulated) distributions of the different likelihood ratios. The distributions analyzed were those corresponding to the maximum LR at each scan or at each chromosome partition. They are not far apart from the theoretical asymptotic values. There is a trend, as expected, in increasing the mean and variance with the degrees of freedom and, in fact, the empirical threshold is sometimes less conservative than the theoretical chi-square figure $P(\chi^2 > x_{0.05}) > 0.05$. Figure 2 shows the empirical cumulative distribution functions (CDFs) together with their chi-square counterparts. We can conclude as Knott and Haley (1992) that, for all practical purposes, the chi-square distribution is a valid approximation in this instance.

Scenario 1: Here there is only one QTL in the linkage group studied. The average LRs over segments are in Figures 3–5 for cases a, b, and c, respectively. These figures are equivalent to a LOD score or *F*-graphics in a chromosome scan, but we prefer a bar representation to underline that they are tests at discrete positions. Note again that the $LR_{0,SM}$ corresponds to a test where the whole chromosome is considered; it changes only the partition employed (Figure 1b). In the presence of a single QTL, the segment mapping test shows a distinct behavior from that of the point scan strategies (mixed, random, and fixed models). As expected, the scan strategies produce LR_0 maxima at the QTL position, and LR_0 decreases as the test position moves away. In contrast, $LR_{0,SM}$ also shows a clear maximum with partition 1, whereas the rest of the partitions show a rather flat and nonclearly decreasing profile. The differences between partitions should be due to random fluctuations because no clear pattern emerges. Now consider LR_s. This statistic should be larger than zero whenever there is a QTL in the position considered and close to zero elsewhere. This is what we observe, and LR_s shows clear maxima at the QTL positions irrespective of the genetic case, a, b, or c. The drop in LR_s when we move away from the QTL position is much larger than in scan methods; *e.g.*, compare the change in $LR_{0,MM}$ and in LR_S between positions 1 and 2 (Figures 3–5).

Although there are some similarities between $LR_{0,MM}$,

TABLE 1

Empirical likelihood-ratio distributions

LR[a]	d.f.[b]	μ[c]	σ^2[d]	x 0.05[e]	$P(\chi^2 > x_{0.05})$[f]
Segment mapping	4	3.30	4.71	7.72	0.102
RS[a]	2	2.54	3.24	6.42	0.040
Mixed model	2	2.03	2.51	5.03	0.081
Random model	1	0.53	1.11	2.92	0.088
Fixed model	1	1.78	1.98	4.55	0.033

[a] LR_s in the segment mapping approach.
[b] Expected degrees of freedom.
[c] Empirical mean obtained by simulation.
[d] Empirical variance.
[e] Empirical 5% significance threshold.
[f] Probability corresponding to a chi square with degrees of freedom.

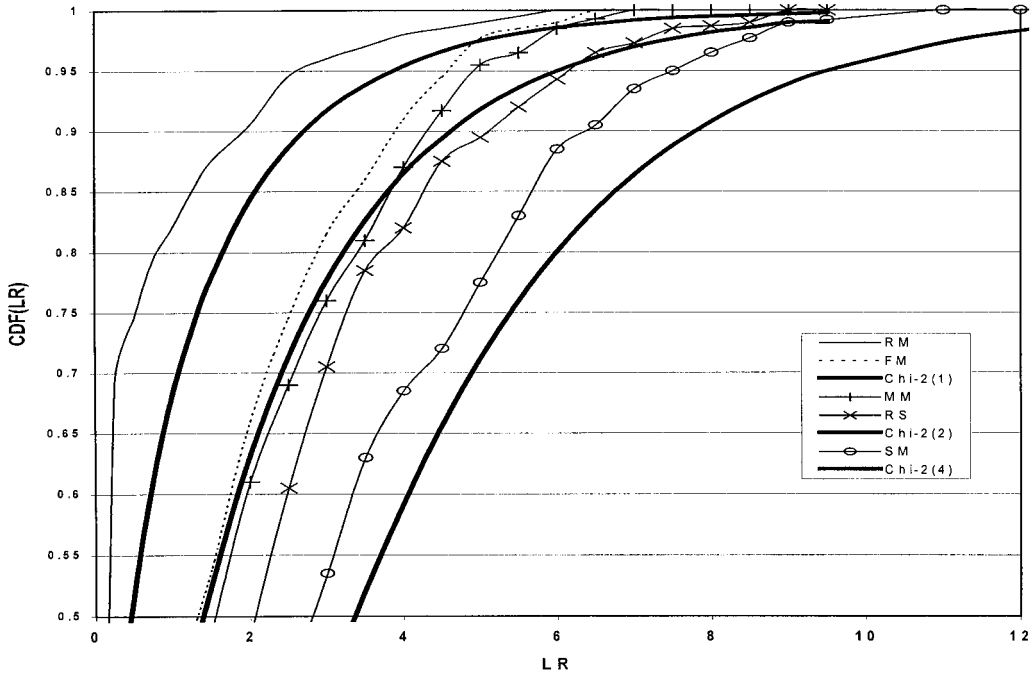

Figure 2.—Empirical and theoretical (Chi-2) cumulative distributions of the several likelihood ratios used in this work. RS corresponds to LR_S in the segment mapping approach; the remaining figures correspond to LR_0 (see text): SM, segment mapping; MM, mixed model; RM, random model; FM, fixed model. The Chi-2 are the solid thick lines, with degrees of freedom in parentheses in the inset.

$LR_{0,RM}$, and $LR_{0,FM}$, their performance depends critically on the underlying genetic model. Consider first case a (Figure 3), where the random model is the most appropriate strategy. It is not surprising that $LR_{0,RM}$ is very close to $LR_{0,MM}$ and $LR_{0,SM}$ in position 1, despite the larger number of parameters involved in the latter two methods. Moreover, Table 2 shows that segment mapping as well as the mixed and random models lead to the same σ^2_s estimate. Segment mapping and the mixed model clearly show that the mean of allelic effects ($\Delta/2$) is zero and that there is no additional variation out of segment 1. All three methods had a 100% power in detecting the QTL. In contrast, the fixed model was the worst strategy considered; not only in 61% of the replicates did maximum LR_0 coincide with the QTL position, but also in only 68% out of those 61% replicates were the $LR_{0,FM}$ significant. The σ^2_e estimate was clearly biased (Table 2).

In contrast, the fixed model (8) is the best choice in scenario 1b because the premise that the QTL affecting the trait are diallelic with alternative alleles fixed in each parental line is fulfilled. Here $LR_{0,RM}$ was much lower

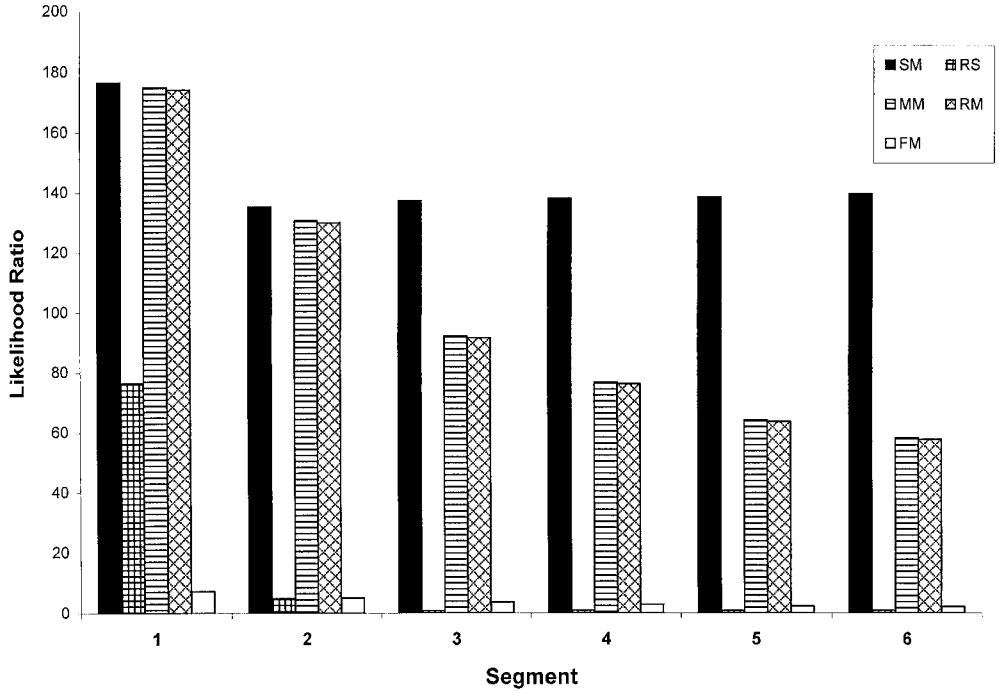

Figure 3.—Bar profiles of the different likelihood ratios at the positions (partitions) considered. RS corresponds to LR_S in the segment mapping approach; the remaining figures correspond to LR_0 (see text): SM, segment mapping; MM, mixed model; RM, random model; FM, fixed model. Scenario 1a (1 QTL, $\sigma^2_A = \sigma^2_B = \sigma^2_e = 1$, $\Delta = 0$).

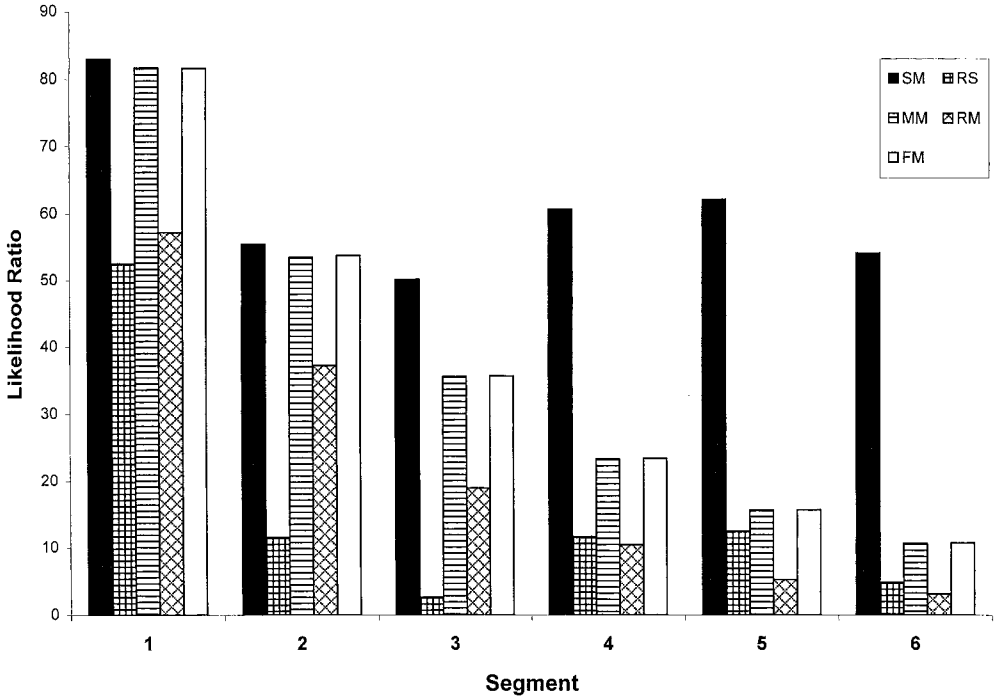

Figure 4.—Bar profiles of the different likelihood ratios at the positions (partitions) considered. RS corresponds to LR_S in the segment mapping approach; the remaining figures correspond to LR_0 (see text): SM, segment mapping; MM, mixed model; RM, random model; FM, fixed model. Scenario 1b (1 QTL, $\sigma^2_A = \sigma^2_B = 0$; $\sigma^2_e = 2$, $\Delta/2 = 1$).

than $LR_{0,FM}$, and this was very similar to $LR_{0,MM}$ and $LR_{0,SM}$, because no additional parameters are needed. The fixed model yielded unbiased estimates of σ^2_e, μ, and $\Delta_s/2$, as did the segment mapping and the mixed-model analysis. A random-model analysis also yielded with power 100% the first position as the most likely one to contain a QTL. But note that total variance ($\sigma^2_e + \sigma^2_s$) was overestimated and the mean estimate was biased downward because the assumed genetic model was not adequate.

The most complex, and realistic, scenario is when alleles are not fixed and their average effect differs from line to line (case c). All four analysis strategies identified the correct QTL location (except for one replicate in the fixed-model analysis) with power 100%, and in this sense all methods would lead to the detection of a QTL. But classical methods, either fixed or random models, are not capable of extracting all available information from the data. According to previous results, it is not surprising that the fixed-model analysis resulted in a biased estimate of σ^2_e, whereas the estimates of Δ were

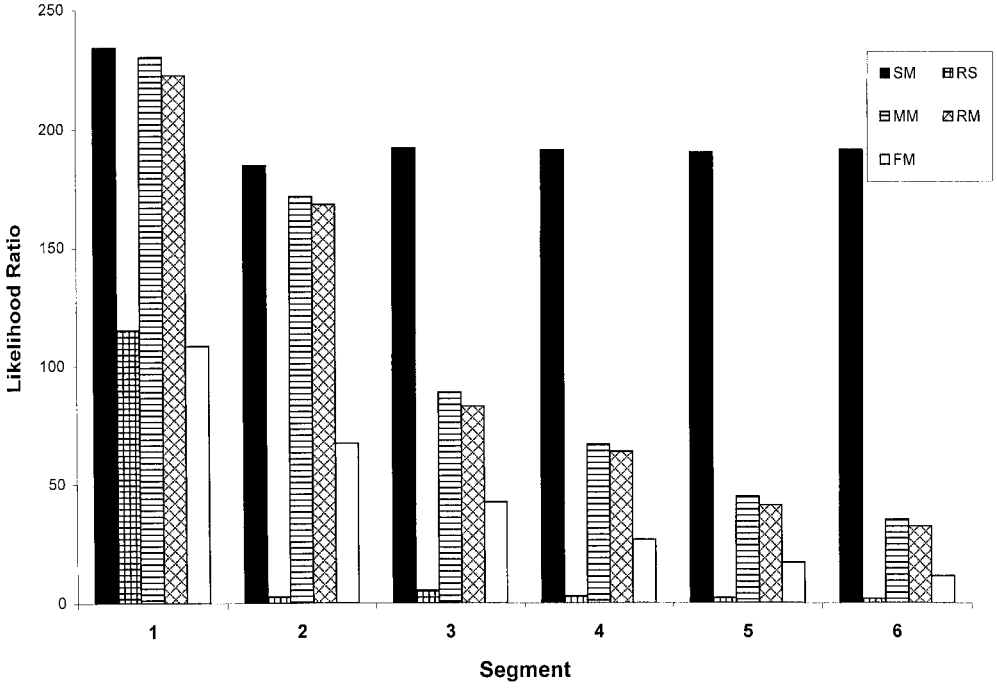

Figure 5.—Bar profiles of the different likelihood ratios at the positions (partitions) considered. RS corresponds to LR_S in the segment mapping approach; the remaining figures correspond to LR_0 (see text): SM, segment mapping; MM, mixed model; RM, random model; FM, fixed model. Scenario 1c (1 QTL, $\sigma^2_A = \sigma^2_B = \sigma^2_e = \Delta/2 = 1$).

TABLE 2

Results with genetic scenario 1 at segment 1 (1–10 cM)

Case	Method	%[a]	Π_1[b]	σ^2_e[c]	$\sigma^2_{\bar{s}}$[d]	σ^2_s[e]	μ[f]	$\Delta_{\bar{s}}/2$[g]	$\Delta_s/2$[h]
a	Segment mapping	100	100	1.02 ± 0.02	0.00 ± 0.00	0.58 ± 0.02	0.06 ± 0.05	0.03 ± 0.02	−0.02 ± 0.03
	Mixed model	100	100	1.02 ± 0.02	—	0.58 ± 0.02	0.06 ± 0.05	—	0.00 ± 0.03
	Random model	100	100	1.02 ± 0.02	—	0.58 ± 0.02	0.06 ± 0.05	—	—
	Fixed model	61	68	1.84 ± 0.07	—	—	0.06 ± 0.06	—	−0.06 ± 0.04
b	Segment mapping	100	100	2.04 ± 0.02	0.00 ± 0.00	0.01 ± 0.01	−0.01 ± 0.01	−0.20 ± 0.03	1.18 ± 0.03
	Mixed model	100	100	2.06 ± 0.02	—	0.00 ± 0.00	−0.02 ± 0.01	—	1.04 ± 0.02
	Random model	100	100	2.03 ± 0.02	—	0.52 ± 0.05	−0.39 ± 0.02	—	—
	Fixed model	100	100	2.06 ± 0.02	—	—	−0.02 ± 0.01	—	1.04 ± 0.02
c	Segment mapping	100	100	0.98 ± 0.02	0.01 ± 0.01	0.59 ± 0.05	−0.04 ± 0.08	−0.22 ± 0.03	1.12 ± 0.09
	Mixed model	100	100	0.98 ± 0.02	—	0.66 ± 0.05	−0.04 ± 0.08	—	0.97 ± 0.08
	Random model	100	100	0.97 ± 0.02	—	0.92 ± 0.07	−0.46 ± 0.07	—	—
	Fixed model	96	100	1.55 ± 0.06	—	—	0.01 ± 0.09	—	0.99 ± 0.09

The statistics are the average of 30 replicates. The average simulated mean QTL variance and $\Delta/2$ were 0.92 and 0.01 in case a and 0.78 and 0.92 in case c, respectively.

[a] Percentage of replicates where partition or segment 1 (the QTL position) corresponded to maximum likelihood.

[b] Power, computed as the percentage of replicates where LR exceeded the empirical 5% significance threshold (see Table 1), out of those where the maximum LR was at position 1.

[c] Residual variance estimate.

[d] Estimate of the genetic variance due to the complement of segment 1 (11–60 cM).

[e] Estimate of the genetic variance due to segment 1.

[f] Estimate of the general mean.

[g] Estimate of the mean difference due to the complement of segment 1.

[h] Estimate of the mean difference due to segment 1.

much more accurate. Alternatively, the RM analysis provided aberrant estimates of the general mean, but σ^2_e and σ^2_s estimates were more realistic. Finally, the mixed model is the most parsimonious and correct model and results in the best estimates. The segment mapping indicates that there is a single segment contributing to the F_2 genetic differences, as can be inferred from the dramatic drop in LR_s for $s > 1$. The estimates of σ^2_s and Δ_s show that the QTL affects both the variance and the mean.

Scenario 2c: Consider first the behavior of the likelihood ratio under the different models of analyses (Figure 6). The scan approaches (mixed, random, and fixed models) peaked at both QTL positions with probability close to 50% in all methods (Table 3) because the two QTL were of about the same effect. Again, $LR_{0,RM}$ was higher than $LR_{0,FM}$, and the power was slightly larger with the random model than with the fixed-model approach. The $LR_{0,SM}$ peaks were more scattered, but almost 50% of the maxima were located at intermediate positions (partitions 3 and 5). These partitions correspond to those where segments containing QTL are grouped *vs.* segments without QTL. We can think of these partitions as the most "reasonable" ones. Occasionally the $LR_{0,SM}$ peaked at partitions 1 or 6 because in that particular replicate a given QTL effect was much larger than the other QTL effect. In no replicate did the maximum $LR_{0,SM}$ coincide with partition 2 or 5. The plot of LR_s clearly indicates that only segments 1 and 6 contain QTL (Figure 6). Moreover, Table 3 shows that SM resulted in unbiased estimates of σ^2_e irrespective of the partition because the variation along the whole chromosome is always considered (at the expense of logically increasing the number of parameters). The other strategies, the mixed and random model, but especially the fixed model, overestimated σ^2_e. The mixed-model point estimates of σ^2_s and of $\Delta/2$ collected the variation along the whole chromosome and not only on that position (a phenomenon already described by Jansen 1993 and Zeng 1993 for the fixed model but we can see that applies equally to the random or mixed models). The random and fixed models provided much poorer estimates than the mixed model.

Scenario 3c: Here the marker positions coincided with segment bounds. The presence of a close but distinct cluster of genes results in a different LR_0 pattern as compared to scenario 2c. The $LR_{0,MM}$ and $LR_{0,RM}$ tend now to peak in between both clusters, whereas $LR_{0,FM}$ results in a completely flat profile, with maxima randomly located along the chromosome (Figure 7, Table 4). The LR_s allows us to identify convincingly that the intermediate segment contains no QTL. Note that LR_s for $s = 1$ and 3 are significant despite the much lower value compared to the other LR. Again $LR_{0,SM}$ peaked at partition 2. The phenomena already described in scenario 2c are noted again but to a larger extent because more than one linked loci are involved now: there is a bias in σ^2_e estimates and the point genetic variance collects the variance from the whole linkage group. Note, *e.g.*, that σ^2_s estimates are the same for all $s = 1$,

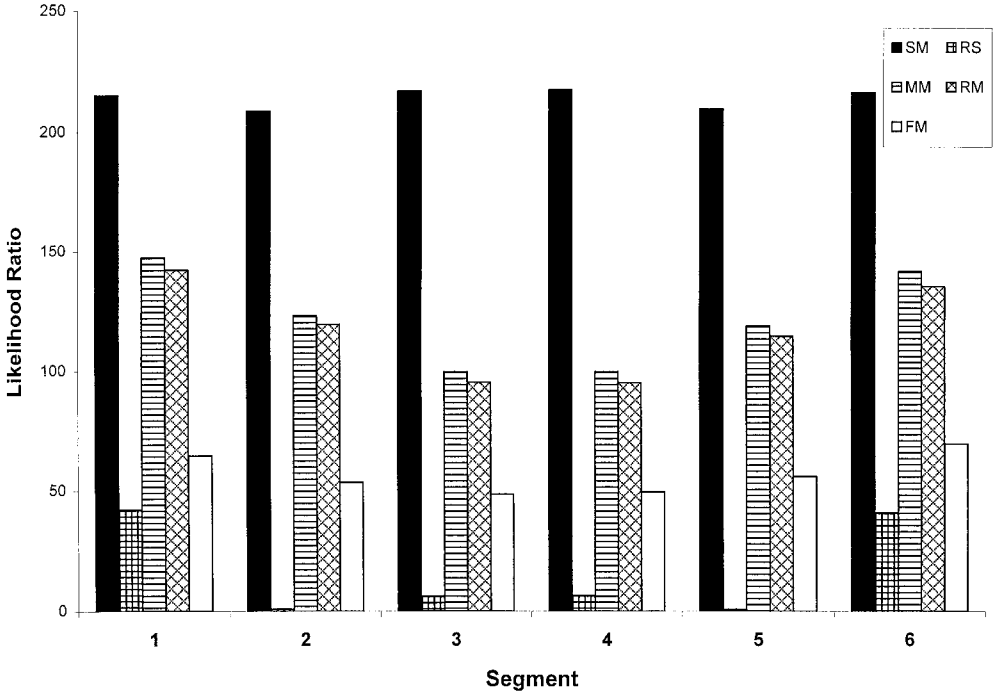

Figure 6.—Bar profiles of the different likelihood ratios at the positions (partitions) considered. RS corresponds to LR_S in the segment mapping approach; the remaining figures correspond to LR_0 (see text): SM, segment mapping; MM, mixed model; RM, random model; FM, fixed model. Scenario 2c (2 QTL, $\sigma^2_A = \sigma^2_B = \sigma^2_e = \Delta/2 = 1$).

3 with the mixed- and random-model analyses, although there are no QTL on positions 20–40 cM. Again it is not surprising that the fixed model provided unrealistic estimates of σ^2_e, whereas the QTL effect estimates (Δ) are confounded, as in scenario 2c. Segment mapping is the most appropriate analysis tool here and it is the only method providing accurate results.

DISCUSSION

The QTL mixed model developed here is a generalization over the Wang *et al.* (1998) approach by allowing that loci can be linked and making use of the information provided by any number of molecular markers jointly; thus the method can be applied to the analysis of QTL studies of F_2 crosses. The methodology presented here shows as well that the covariance between F_2 individuals should be split into the probabilities of identity by descent contributed by each breed. Further, the segment-mapping approach allows a global analysis by partitioning the genome, or the chromosome, in segments. Rodolphe and Lefort (1993) proposed considering the whole genome simultaneously but their approach is a fixed model with multiple regression on all markers genotyped. And this results in a loss of power as the number of markers increases. This does not occur with segment mapping because the number of parameters depends on the number of segments defined, not on the number of markers used.

The simulation results presented show that, under a variety of genetic architectures, the mixed-model and segment-mapping procedures are more robust and flexible strategies than the classical methods based on pure fixed or random models. Segment-mapping, mixed model, and pure fixed or random models are hierarchical levels of analysis complexity, as can be seen from comparing (5), (6), (7), and (8). A likelihood-ratio test can be used to decide whether there is evidence to consider a genetic model more complex than the one assumed in classical methods. Overall, the point mixed model showed optimum performance with a single QTL. The segment-mapping approach will be most useful in the case of linked QTL (Tables 3 and 4). The LR_s will help to determine which chromosome regions are likely to contain QTL. It is interesting that the segment-mapping partition corresponding to the maximum likelihood (at equal number of parameters) occurs when the genome is partitioned according to its effect on the trait. For instance, when the QTL are in both extremes, the likelihood is maximized when a model-partitioning segment equidistant between the two QTL or the two clusters *vs.* the rest of the genome is chosen (Tables 3 and 4). But it is also a nice property of segment mapping that, irrespective of the partition actually chosen, it results in general in accurate estimates of σ^2_e and of the total contribution of the chromosome, $\sigma^2_s + \sigma^2_{\bar{s}}$ and $\Delta_s + \Delta_{\bar{s}}$. This contrasts with fixed-, random-, or mixed-model approaches, where accurate estimates are obtained only at the exact position of the QTL.

The classical fixed-model approach is simple to compute and easy to interpret in F_2 crosses, although it makes very strong assumptions about allele distributions in the parental lines. We have shown that fixed-model estimates can be dramatically affected if alleles are not fixed within lines, even in one-locus scenarios (Tables

TABLE 3

Results with genetic scenario 2c

Method	s^a	$\%^b$	$\Pi_1{}^c$	$\sigma_e^{2\,d}$	$\sigma_s^{2\,e}$	$\sigma_s^{2\,f}$	μ^g	$\Delta_{\bar{s}}/2^h$	$\Delta_s/2^i$
Segment mapping	1	30	100	1.02 ± 0.02	0.77 ± 0.07	0.37 ± 0.04	0.06 ± 0.07	0.43 ± 0.09	0.64 ± 0.08
	2–5	47	100	1.03 ± 0.02	1.14 ± 0.08	0.00 ± 0.00	0.05 ± 0.07	1.50 ± 0.13	0.43 ± 0.07
	6	23	100	1.03 ± 0.02	0.79 ± 0.07	0.32 ± 0.05	0.03 ± 0.08	0.44 ± 0.08	0.62 ± 0.10
Mixed model	1	53	100	1.35 ± 0.03	—	0.67 ± 0.06	0.07 ± 0.08	—	0.79 ± 0.06
	2–5	0	—	1.47 ± 0.04	—	0.84 ± 0.10	0.02 ± 0.07	—	0.74 ± 0.06
	6	47	100	1.37 ± 0.04	—	0.70 ± 0.07	0.00 ± 0.08	—	0.81 ± 0.08
Random model	1	50	100	1.34 ± 0.03	—	0.87 ± 0.07	−0.26 + 0.07	—	—
	2–5	0	—	1.47 ± 0.04	—	1.03 ± 0.10	−0.28 ± 0.06	—	—
	6	50	100	1.37 ± 0.04	—	0.94 ± 0.07	−0.33 ± 0.06	—	—
Fixed model	1	43	85	1.86 ± 0.06	—	—	0.00 ± 0.07	—	0.79 ± 0.07
	2–5	0	—	1.92 ± 0.07	—	—	0.00 ± 0.07	—	0.72 ± 0.06
	6	57	100	1.85 ± 0.07	—	—	0.00 ± 0.07	—	0.79 ± 0.08

The statistics are the average of 30 replicates. The average simulated mean QTL variance and $\Delta/2$ were 0.48 and 0.54 for the first QTL and 0.42 and 0.55 for the second QTL.

[a] Segment (partition) order.

[b] Percentage of replicates where partition or segment 1 (the QTL position) corresponded to maximum likelihood.

[c] Power, computed as the percentage of replicates where LR exceeded the empirical 5% significance threshold (see Table 1), out of those where the maximum LR was at position 1.

[d] Residual variance estimate.

[e] Estimate of the genetic variance due to the complement of segment 1 (11–60 cM).

[f] Estimate of the genetic variance due to segment 1.

[g] Estimate of the general mean.

[h] Estimate of the mean difference due to the complement of segment 1.

[i] Estimate of the mean difference due to segment 1.

2 and 4). A systematic upward bias of the σ_e^2 estimate was observed in particular. Allele segregation also results in a loss of power with the fixed model (Alfonso and Haley 1998), and it can be seen that the $LR_{0,FM}$ is lower in case a and c than in b, when alleles are fixed (Table 2). In contrast, segment mapping gave reasonable estimates of the QTL mean effects and variance. All in all, it cannot be overlooked that the standard regression approach (Lander and Botstein 1989; Haley and Knott 1992) has been successful in identifying QTL in crosses between outbred lines. Some of these QTL have been confirmed in independent experiments (*e.g.*, Andersson *et al.* 1994; Walling *et al.* 1998; M. Perez-Enciso, A. Clop, J. L. Noguera, C. Óvilo, A. Coll, J. Fulch, D. Babot, J. Estany, M. A. Oliver, I. Diaz and A. Sánchez, unpublished results, for a QTL on chromosome 4 affecting fatness in pigs), strongly suggesting that they are not false positives and that allele effects are distinct between breeds. Note (Table 2) that the fixed model will tend to identify the correct QTL position even if all genetic assumptions are not fulfilled, at the price of biased estimates and misleading significance levels. The fixed model can be generalized to deal with more than one QTL using cofactors or an n-QTL model, but the presence of gene clusters inevitably causes individual QTL not to be resolved individually, and estimates obtained with a genome scan approach will probably be unreliable. In addition, more than one QTL worsens the performance of the fixed model if the alleles are not fixed within breeds.

It is interesting to compare the performance of random and fixed models under the genetic models considered. The random model was more robust than the fixed-model approach in terms of locating a QTL: the $LR_{0,RM}$ was higher in case b (Figure 4) than $LR_{0,FM}$ in case a (Figure 3), as well as in case c (Figures 5, 6, and 7). That is, the random model behaved better when the random-model assumptions were violated than the fixed model did when fixed-model assumptions did not hold. This is an interesting result; the random model does not seem *a priori* a reasonable strategy for analyzing F_2 crosses as no differences in allelic effects between breeds are assumed. Xu (1998) studied by computer simulation the performance of random models in analyzing crosses but in a context where several crosses between different inbred lines were analyzed together. We are not aware of actual F_2 QTL experiments analyzed using a completely random model. Nonetheless De Koning *et al.* (1999) have analyzed a F_2 cross in pigs using a within-sire regression approach (Knott *et al.* 1996) and a classical fixed model. The former method does not make specific assumptions about number of alleles and frequencies in the parental lines, at the expense of increasing the number of parameters and disregarding genotypic information of dam origin. Interestingly, the two statistical approaches lead to distinct results, both in QTL effect and

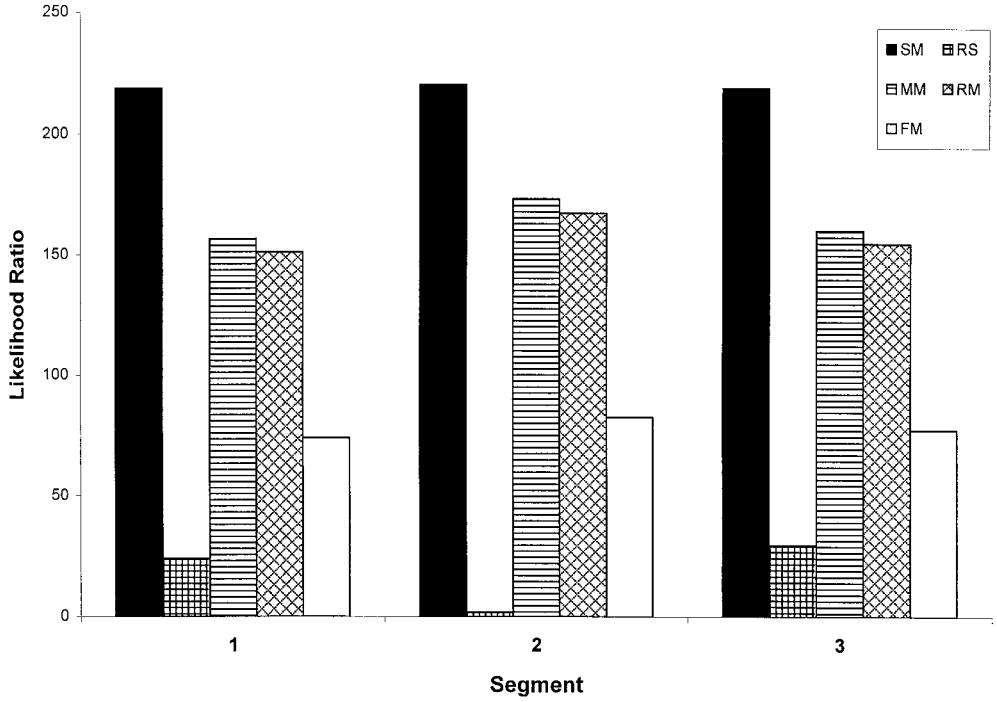

Figure 7.—Bar profiles of the different likelihood ratios at the positions (partitions) considered. RS corresponds to LR_S in the segment mapping approach; the remaining figures correspond to LR_0 (see main text): SM, segment mapping; MM, mixed model; RM, random model; FM, fixed model. Scenario 3c (40 QTL, $\sigma^2_A = \sigma^2_B = \sigma^2_e = \Delta/2 = 1$).

in location (with the exception of a QTL for backfat thickness on chromosome 7). The within-sire approach exhibited, overall, smaller power than the fixed model. This analysis seems to contradict our simulation results concerning the robustness of the random model, but there are important differences between the random model and the within-sire regression. First, the within-sire regression as used by De Koning *et al.* (1999) disregards dam information. This can have a negligible effect in very large and outbred populations, but not necessarily so in modest family sizes (22–51 half-sibs in De Koning *et al.* 1999) and in a F_2 between divergent breeds where the variation contributed by the meiotic segregation in the dam can be large compared to the environmental variance. Second, we have assumed in the simulations a maximum informativity in terms of marker alleles, and it is plausible that the relative performance of the methods differs at lower levels of heterozygosity.

The approximation of (3) depends on the informativity and density of molecular markers. We have not explored in detail the impact of noninformativeness on the segment mapping approach, but it can be seen that the partitions used in genetic scenarios 1 and 2 (Tables 2 and 3) have segments with one bound not coinciding with markers, *i.e.*, the least informative possible situation. Despite this, the estimates were quite reasonable. Take, *e.g.*, genetic scenario 1 (Table 2): in partition 1 the variance associated with segment 1–10 cM collects almost all genetic variance and σ^2_s is zero, as it should be. In scenario 2c the only partitions where the 10-cM segment collects a significant variance are the first and last, where QTL are actually located (Table 3). In addition, the LR_s statistic has a very distinct behavior depending on whether or not there is a QTL in the particular segment under consideration (Figures 3–7).

The simulations carried out here have assumed that loci behave additively, both between and within breeds. This may seem a quite strong assumption in view of the ample empirical evidence for heterosis in line crosses (Lynch and Walsh 1998). The general theory to deal with dominance in crosses between outbred lines has been developed by Lo *et al.* (1995), and it can be extended to deal with molecular markers. Unfortunately the number of parameters that need to be estimated is very large so that in practice one may be confined to providing only approximate estimates of the dominance variance or making strong assumptions about allele distributions. The fixed-model approach and regression-type methods take into account dominance by adding an additional covariable to the probability of the QTL being heterozygous at the position of interest. The same course of action can be followed here, but it should be noted that this strategy presupposes that a diallelic locus is fixed in each line. Otherwise, the dominance deviation estimate will be biased and not accurate.

We have assumed a model $\sigma^2_A = \sigma^2_B$, *i.e.*, equal genetic variances across the parental lines, in the analyses reported here. Note, however, that the theory developed allows us to distinguish between genetic variances in each breed. To test this, we ran 30 additional replicates in scenario 1 with parameters $\sigma^2_e = \sigma^2_A = 1$ and $\Delta = \sigma^2_B = 0$. We analyzed the data using a random model with σ^2_A and σ^2_B as distinct parameters. The average actual simulated value for σ^2_A was 0.901, and the estimates were 0.98 ± 0.02 (σ^2_e), 0.90 ± 0.07 (σ^2_A), and 0.01 ± 0.00 (σ^2_B). The estimate of σ^2_B was exactly 0 in 14 replicates.

TABLE 4

Results with genetic scenario 3c

Method	s^a	%^{b}	$\Pi_1{}^c$	$\sigma^{2\,d}_e$	$\sigma^{2\,e}_s$	$\sigma^{2\,f}_s$	μ^g	$\Delta_{\bar{s}}/2^h$	$\Delta_s/2^i$
Segment mapping	1	30	100	1.04 ± 0.02	0.53 ± 0.04	0.29 ± 0.04	−0.03 ± 0.10	0.60 ± 0.08	0.44 ± 0.07
	2	57	100	1.03 ± 0.01	0.72 ± 0.06	0.06 ± 0.02	−0.02 ± 0.10	1.24 ± 0.09	−0.16 ± 0.07
	3	17	100	1.04 ± 0.01	0.50 ± 0.05	0.31 ± 0.04	−0.03 ± 0.10	0.59 ± 0.08	0.45 ± 0.07
Mixed model	1	23	100	1.28 ± 0.03	—	0.67 ± 0.07	−0.02 ± 0.10	—	0.82 ± 0.07
	2	40	100	1.22 ± 0.02	—	0.72 ± 0.05	−0.08 ± 0.11	—	0.85 ± 0.06
	3	37	100	1.28 ± 0.04	—	0.68 ± 0.05	−0.03 ± 0.11	—	0.85 ± 0.06
Random model	1	23	100	1.27 ± 0.03	—	0.92 ± 0.07	−0.38 ± 0.11	—	—
	2	40	100	1.22 ± 0.03	—	0.92 ± 0.05	−0.44 ± 0.10	—	—
	3	37	100	1.27 ± 0.04	—	0.89 ± 0.05	−0.38 ± 0.10	—	—
Fixed model	1	33	100	1.80 ± 0.07	—	—	−0.07 ± 0.10	—	0.88 ± 0.07
	2	33	90	1.77 ± 0.07	—	—	−0.07 ± 0.10	—	0.92 ± 0.08
	3	33	100	1.80 ± 0.08	—	—	−0.07 ± 0.10	—	0.89 ± 0.07

The statistics are the average of 30 replicates. The average simulated mean QTL variance and $\Delta/2$ were 0.41 and 0.58 for the first cluster, and 0.42 and 0.46 for the second cluster.

[a] Segment (partition) order.

[b] Percentage of replicates where partition or segment 1 (the QTL position) corresponded to maximum likelihood.

[c] Power, computed as the percentage of replicates where LR exceeded the empirical 5% significance threshold (see Table 1), out of those where the maximum LR was at position 1.

[d] Residual variance estimate.

[e] Estimate of the genetic variance due to the complement of segment 1 (11–60 cM).

[f] Estimate of the genetic variance due to segment 1.

[g] Estimate of the general mean.

[h] Estimate of the mean difference due to the complement of segment 1.

[i] Estimate of the mean difference due to segment 1.

A likelihood ratio showed that a model including σ^2_B did not improve over a model without σ^2_B. The approach developed here thus provides insight into the genetic architecture of the trait in the parental lines, as it should allow us to estimate $\sigma^2_{A,s}$ and $\sigma^2_{B,s}$ for each segment considered. These are the most relevant parameters in the study of an outbred population and it is a bonus of the usefulness of F_2 crosses. With current statistical approaches, the only loci detected with maximum power are those with alleles fixed within line, which limits the inferences with respect to loci segregating in the parental lines. Moreover, the mixed model and segment mapping encourage the use of performance records from F_1 and parental individuals not usually analyzed jointly with F_2 records nor even recorded. F_1 and parental records can be analyzed jointly with the F_2 data without any significant modification of (1)–(4). An advantage of including these records is that they will provide insight into the presence and extent of dominance action.

Note that in segment mapping we do not make the distinction between a QTL and a polygenic background, and it is not necessarily assumed in segment mapping that a single locus is segregating within the segment or segments considered. It follows that it is more relevant in the segment-mapping context to test whether a given segment, however small, contributes significantly to genetic variation than in an accurate QTL location, as is emphasized in interval mapping (*e.g.*, Visscher *et al.* 1996). The importance of accuracy of QTL location or correctly ascertaining the number of QTL need not be overestimated. First, if a very dense genotyping is carried out, segment-mapping will be able to separate intervals contributing to variation more effectively than genome scan because external "genetic noise" is properly accounted for in segment mapping. Compare, *e.g.*, the drops in $LR_{0,MM}$ and LR_s between positions 1 and 2, which have very similar distributions under the null hypothesis (Figure 2). The change in LR_s is larger than in $LR_{0,MM}$ for all genetic cases. We may thus conjecture that a combination of $LR_{0,SM}$ and LR_s tests may lead to a more accurate location of the QTL than a simple scan with $LR_{0,MM}$, although more extensive simulation is needed to prove this. Second, the candidate genes will be readily located once a promising region is identified as genetic maps are becoming densely populated with known genes. The current strategy in QTL analysis is to look for candidate genes within the chromosome regions that have shown association with the trait. It is likely, in fact, that the reverse strategy will be predominant in the future: once the number of cloned candidate genes becomes very large and their physiological effects are ascertained or inferred, it will be routine to estimate the fraction of genetic variance associated with these genes, including possible epistatic effects, in a particular population.

A feature of the segment-mapping strategy is that there is not an obvious course of action to conduct a

genome partitioning. We propose to run a preliminary analysis with a segment partitioning scan as depicted in Figure 1b complemented with LR_s tests every, say, 10 or 5 cM. This should allow us to identify which segments are more promising. In a second analysis the noninteresting regions should be discarded from further consideration, and a detailed partitioning of the most relevant genome regions can be studied, together with elucidating whether fixed, random, or mixed models are more suitable for each segment. Interactions between segments can be analyzed as well. The ultimate goal of segment mapping would be to have a function establishing the appropriate weights given to each region of the genome when computing the additive relationship between animals and, additionally, the expected changes in mean as well. Given estimation errors, the most parsimonious model explaining the maximum variance should be chosen. A reasonable compromise is to classify genome regions according to their effect on the trait of interest, *e.g.*, strong, weak, and nonsignificant. Regions of similar effect can be analyzed together in the same segment. Note that different "segmentation" may be used to model variance components or means; *i.e.*, the whole genome may be partitioned in just three segments grouped according to its contribution to total genetic variance, whereas differences in means (Δ_s) can be fitted in more segments, or at specific genome locations if there is clear evidence of a QTL. In that manner, it can be considered that QTL that contribute to differences between lines do not contribute necessarily to differences within lines.

In conclusion, we have put forward a methodology based on mixed-model theory that allows for complex genetic models and, at least theoretically, a simultaneous analysis of the whole genome. It has been shown that genome scans using regression or completely random model approaches are but particular cases of the theory presented in this work. The random model shows a more robust behavior than the most commonly used regression approach. Finally, segment-mapping principles can be accommodated to a variety of experimental designs, not only F_2 crosses.

We are grateful to Miguel Toro, Luis Silió, Rohan Fernando, and the referees for useful comments. Some of this work was accomplished during a sabbatical visit of M.P.E. to Iowa State University. M.P.E. expresses his appreciation for the financial support received by Cotswold USA and Max Rothschild during his stay at Iowa State University. Work was funded by projects Comisión Asesora de Ciencia y Tecnología AGF96-2510 (Spain) and BIO4-CT97-962243 (E.U.).

LITERATURE CITED

Alfonso, L., and C. S. Haley, 1998 Power of different F_2 schemes for QTL detection in livestock. Anim. Prod. **66:** 1–8.

Andersson, L., C. S. Haley, H. Ellegren, S. A. Knott, M. Johansson *et al.*, 1994 Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science **263:** 1771–1774.

Bovenhuis, H., J. A. M. Van Arendonk, G. Davis, J. M. Elsen, C. S. Haley *et al.*, 1997 Detection and mapping of quantitative trait loci in farm animals. Livest. Prod. Sci. **52:** 135–144.

De Koning, D. J., L. L. G. Janss, A. P. Rattink, P. A. M. Van Oers, B. J. De Vries *et al.*, 1999 Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (Sus scrofa). Genetics **152:** 1679–1690.

Elsen, J. M., B. Mangin, B. Goffinet, D. Boichard and P. Le Roy, 1999 Alternative models for QTL detection in livestock. I. General introduction. Genet. Sel. Evol. **31:** 213–224.

Fernando, R. L., and M. Grossman, 1989 Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21:** 467–477.

Goddard, M. E., 1992 A mixed model for analysis of data on multiple genetic markers. Theor. Appl. Genet. **83:** 878–886.

Goldgar, D. E., 1990 Multiplepoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. **47:** 957–967.

Grattapaglia, D., F. L. G. Bertolucci and R. R. Sederoff, 1995 Genetic mapping of QTLs controlling vegetative propagation in Eucaliptus grandis and E. urophylla using a pesudo-testcross mapping startegy and RAPD markers. Theor. Appl. Genet. **90:** 933–947.

Grignola, F. E., I. Hoeschele and B. Tier, 1996 Mapping quantitative trait loci in outcross populations via residual maximum likelihood. Genet. Sel. Evol. **28:** 479–490.

Haley C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

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.

Heath, S., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Hill, W. G., 1993 Variation in genetic composition in backcrossing programs. J. Hered. **84:** 212–213.

Hoeschele, I., P. Uimari, F. E. Grignola, Q. Zang and K. M. Gage, 1997 Advances in statistical methods to map quantitative trait loci in outbred populations. Genetics **147:** 1445–1457.

Hunt, G. J., E. Guzman-Novoa, M. K. Fondrik and R. E. Page, Jr., 1998 Quantitative trait loci for honey bee stinging behavior and body size. Genetics **148:** 1203–1213.

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

Knott, S. A., and C. S. Haley, 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. **60:** 139–151.

Knott, S. A., J. M. Elsen and C. S. Haley, 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theor. Appl. Genet. **93:** 71–80.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lo, L. L., R. L. Fernando and M. Grossman, 1993 Covariance between relatives in multibreed populations: additive model. Theor. Appl. Genet. **87:** 423–430.

Lo, L. L., R. L. Fernando, R. Cantet and M. Grossman, 1995 Theory for modelling means and covariances in a two-breed population with dominance inheritance. Theor. Appl. Genet. **90:** 49–62.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer, Sunderland, MA.

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

Visscher, P. M., R. Thompson and C. S. Haley, 1996 Confidence intervals in QTL mapping by bootstrapping. Genetics **143:** 1013–1020.

Walling, G., A. L. Archibald, J. A. Cattermole, A. C. Downing, H. A. Finlayson *et al.*, 1998 Mapping of quantitative trait loci on porcine chromosome 4. Anim. Genet. **29:** 415–424.

Wang, T., R. L. Fernando and M. Grossman, 1998 Genetic evaluation by best linear unbiased prediction using marker and trait information in a multibreed population. Genetics **148:** 507–516.

Wright, S., 1968 *Evolution and the Genetics of Populations. Vol. 1. Genetic and Biometric Foundations.* The University of Chicago Press, Chicago.

Xu, S., 1998 Mapping quantitative trait loci using multiple families of linecrosses. Genetics **148:** 517–524.

Xu, S., and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141:** 1189–1197.

Zeng, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

Communicating editor: C. Haley

APPENDIX

The variance/covariance matrix of additive genetic values in the F_2 generation, **G**, is derived. First a finite number of loci (nloci) is considered and then extended to an infinitesimal model. Genetic equilibrium and additive genic action, within and between breeds, is assumed. The genetic value of individual *i* from breed A is

$$g_{\mathrm{A}_i} = \sum_{k=1}^{n_{\mathrm{loci}}} g^{\,\mathrm{S}}_{\mathrm{A}_{i,k}} + \sum_{k=1}^{n_{\mathrm{loci}}} g^{\,\mathrm{D}}_{\mathrm{A}_{i,k}},$$

where $g^{\,\mathrm{S}}_{\mathrm{A}_{i,k}}$ is the sire's origin allele and $g^{\mathrm{D}}_{\mathrm{A}_{i,k}}$ is dam's origin allele at the *k*th locus. Assume for simplicity but without loss of generality that all alleles from all loci are assumed, *a priori*, to have equal effects on the trait. Then

$$g^{\,h}_{\mathrm{A}_{i,k}} \sim N(\mu_k + \Delta_k/2,\ \sigma^2_{\mathrm{A}_k}),$$

$$g^{\,h}_{\mathrm{B}_{i,k}} \sim N(\mu_k - \Delta_k/2,\ \sigma^2_{\mathrm{B}_k}),$$

where

$$\mu_k = \mu/(2n_{\mathrm{loci}}),$$

$$\Delta_k = \Delta/(2n_{\mathrm{loci}}),$$

$$\sigma^2_{\mathrm{A}_k} = \sigma^2_{\mathrm{A}}/(2n_{\mathrm{loci}}),$$

and

$$\sigma^2_{\mathrm{B}_k} = \sigma^2_{\mathrm{B}}/(2\ n_{\mathrm{loci}}), \quad \forall\ h, i, k;$$

h is the haplotype (S or D origin). Breeding values in the F_1 are distributed as $N[\mu, (\sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}})/2]$. The variance of F_2's additive values is given by

$$\mathrm{Var}(g_i) = \mathrm{Var}\left(\sum_{h=1}^{2}\sum_{k=1}^{n_{\mathrm{loci}}} g^h_{i,k}\right) = \sum_{h'=1}^{2}\sum_{h=1}^{2}\mathrm{Cov}\left(\sum_{k=1}^{n_{\mathrm{loci}}} g^h_{i,k},\ \sum_{k=1}^{n_{\mathrm{loci}}} g^{h'}_{i,k}\right),$$

and provided the individual is not inbred,

$$\mathrm{Var}(g_i) = \sum_{h=1}^{2}\mathrm{Var}\left(\sum_{k=1}^{n_{\mathrm{loci}}} g^h_{i,k}\right) = \sum_{h=1}^{2}\sum_{k=1}^{n_{\mathrm{loci}}}\sum_{k=1}^{n_{\mathrm{loci}}}\mathrm{Cov}(g^h_{i,k},\ g^h_{i,k'}).$$

Define as in Lo *et al.* (1993) a variable $w_{k,k'}$ that takes values *AA*, *AB*, *BA*, and *BB* according to the breed origin of each allele at loci *k* and *k*′:

$$\mathrm{Var}(g_i) = \sum_{h=1}^{2}\sum_{k=1}^{n_{\mathrm{loci}}}\sum_{k'=1}^{n_{\mathrm{loci}}}\left\{\mathop{\mathbf{E}}_{w}[\mathrm{Cov}(g^h_{i,k},\ g^h_{i,k'}|w_{kk'})] + \mathop{\mathrm{Cov}}_{w}[\mathbf{E}(g^h_{i,k}|w_{kk'}),\ \mathbf{E}(g^h_{i,k'}|w_{kk'})]\right\}. \tag{A1}$$

The first term in (A1), $\mathrm{Cov}(g^h_{i,k},\ g^h_{i,k'}|w_{k,k'})$, is zero if $k \neq k'$ because linkage equilibrium is assumed within pure breeds or if $w_{k,k'} = AB$ or $w_{k,k'} = BA$. For $k = k'$ it is $\sigma^2_{\mathrm{A}_k}$ or $\sigma^2_{\mathrm{B}_k}$ depending on the origin of *k* (*A* or *B*). Thus,

$$\sum_{h=1}^{2}\sum_{k=1}^{n_{\mathrm{loci}}}\sum_{k=1}^{n_{\mathrm{loci}}}\mathbf{E}[\mathrm{Cov}(g^h_{i,k},\ g^h_{i,k'}\mid w_{kk'})] = p_i\,\sigma^2_{\mathrm{A}} + (1 - pi)\sigma^2_{\mathrm{B}}, \tag{A2}$$

where p_i is the fraction of the genome of origin *A*. The second term in (A1) is

$$\begin{aligned}\mathop{\mathrm{Cov}}_{w}[\mathbf{E}(g^h_{i,k}\mid w_{kk'}),\ \mathbf{E}(g^h_{i,k'}\mid w_{kk'})] &= \mathop{\mathbf{E}}_{w}[\mathbf{E}(g^h_{i,k}\mid w_{kk'})\mathbf{E}(g^h_{i,k}\mid w_{kk'})]\\ &\quad - \mathop{\mathbf{E}}_{w}[\mathbf{E}(g^h_{i,k}\mid w_{kk'})]\mathop{\mathbf{E}}_{w}[\mathbf{E}(g^h_{i,k'}\mid w_{kk'})]\\ &= p^h_i(1 - r_{k,k'})(\mu_k + \Delta_k/2)^2\\ &\quad + (1 - p^h_i)(1 - r_{k,k'})(\mu_k - \Delta_k/2)^2\\ &\quad + r_{k,k'}(\mu^2_k - \Delta^2_k/4)\\ &\quad - [p^h_i(\mu_k + \Delta_k/2)\\ &\qquad + (1 - p^h_i)(\mu_k - \Delta_k/2)]^2,\end{aligned} \tag{A3}$$

where $r_{k,k'}$ is the recombination fraction between loci *k* and *k*′. Combining (A2) and (A3) into (A1) and rearranging,

$$\mathrm{Var}(g_i) = p_i\sigma^2_{\mathrm{A}} + (1 - p_i)\sigma^2_{\mathrm{B}} + \sum_{h=1}^{2}p^h_i(1 - p^h_i)\Delta^2 + \sum_{h=1}^{2}\sum_{k=1}^{n_{\mathrm{loci}}}\sum_{k=1}^{n_{\mathrm{loci}}}[\Delta_k(\mu_k - \Delta_k/2) - 2p^h_i\Delta_k\mu_k]r_{k,k'}. \tag{A4}$$

Setting $r_{k,k'} = 0.5$ for all $k \neq k'$, we retrieve the equation by Lo *et al.* (1993) for an arbitary number of unlinked loci. The last two terms in (A4) are the segregation variance when loci are linked. Equation A4 can be generalized to an infinite number of loci by integrating $r_{k,k'}$ over the whole genome comprising n_{chr} chromosomes of length L_c using results in Hill (1993) for Haldane's mapping function,

$$\mathrm{Var}(g_i) = p_i\sigma^2_{\mathrm{A}} + (1 - p_i)\sigma^2_{\mathrm{B}} + \sum_{c=1}^{n_{\mathrm{chr}}}\left[\sum_{h=1}^{2}p^h_{i,c}(1 - p^h_{i,c})\Delta^2_c + \sum_{h=1}^{2}[\Delta_c(\mu_c - \Delta_c/2) - 2p^h_{i,c}\Delta_c\mu_c]\,\bar{r}\right], \tag{A5}$$

where $\bar{r} = \frac{1}{2} - [2L_c - 1 + \exp(-2\ L_c)]/4\ L^2_c$, when Haldane's mapping function is assumed, *L* is in morgans, $p^h_{i,c}$ is the fraction of chromosome *c*, haplotype *h* of individual *i* of breed origin A, μ_c is the mean effect of loci located in chromosome *c*, and Δ_c is the average difference between loci from each breed origin for chromosome *c*.

In the absence of marker information, $p^h_{i,c}$ is 0.5 along the whole genome and for all F_2 individuals, and (A5) is consequently of little relevance. Now consider that

molecular information such that the probability of breed origin $p_i^h(x)$ can be obtained at any point x of the genome and the genome is partitioned in a series of segments. The genetic variance conditional on marker information is

$$\mathrm{Var}(g_i) = p_i\sigma_A^2 + (1 - p_i)\sigma_B^2 + \sum_{s=1}^{n_{seg}}\left[\sum_{h=1}^{2} p_{i,s}^h(1 - p_{i,s}^h)\Delta_s^2 + \sum_{h=1}^{2}[\Delta_s(\mu_s - \Delta_s/2) - 2p_{i,s}^h\Delta_s\mu_s]\bar{r}\right],$$

where μ_s and Δ_s are the mean of loci in segment s and the average deviation of that particular segment. The null hypothesis is that the contribution to total variation and differences between lines is proportional to genome length, *i.e.*, $\mu_s = \mu L_s/2L$, $\Delta_s = \Delta L_s/2L$, with $L = \Sigma_{s=1}^{n_{seg}} L_s$. Thus,

$$\begin{aligned}\mathrm{Var}(g_i) = {} & p_i\sigma_A^2 + (1 - p_i)\sigma_B^2 \\ & + \Delta^2/(4L^2)\sum_{s=1}^{n_{seg}} L_s^2\left[\sum_{h=1}^{2} p_{i,s}^h(1 - p_{i,s}^h)\right] \\ & + [\Delta(\mu - \Delta/2)]/(2L^2)\sum_{s=1}^{n_{seg}} L_s^2\bar{r}_s \\ & - \Delta\mu/(2L^2)\sum_{s=1}^{n_{seg}}\bar{r}_s\left[\sum_{h=1}^{2}(p_{i,s}^h\,\Delta_s\mu_s)\right]. \qquad \text{(A6)}\end{aligned}$$

The last three terms in (A6) can be neglected: (1) if molecular markers are relatively close, $\bar{r}_s$ and $p_{i,s}^h$ $(1 - p_{i,s}^h)$ tend to zero; (2) the segment's mean breeding value, μ_s, will be negligible in most cases if a general mean is included in model (1); and (3) the sum $\Sigma_{s=1}^{n_{seg}} L_s^2/L^2$ also becomes zero for a large number of small segments. Consequently the diagonal elements of **G** can be simplified as

$$\mathrm{Var}(g_i) \approx p_i\sigma_A^2 + (1 - p_i)\sigma_B^2.$$

In practice one is interested in assessing the particular contribution of a given genome segment, as genetic covariance between individuals is not strictly proportional to the percentage of genome shared; rather, this percentage needs to be weighed by the relevance of each genome location, $\sigma_{A,s}^2$ and $\sigma_{B,s}^2$ in breeds A and B, respectively. Then,

$$\mathrm{Var}(g_i) \approx \sum_{s=1}^{n_{seg}}\sum_{h=1}^{2}[p_{i,s}^h\sigma_{A,s}^2 + (1 - p_{i,s}^h)\sigma_{B,s}^2],$$

with

$$\sigma_A^2 = \sum_{s=1}^{n_{seg}}\sigma_{A,s}^2 \quad \text{and} \quad \sigma_{B,s}^2 = \sum_{s=1}^{n_{seg}}\sigma_{B,s}^2.$$