Quantitative Trait Locus Analysis in Crosses Between Outbred Lines With Dominance and Inbreeding

Miguel Pérez-Enciso,* Rohan L. Fernando,† Jean-Pierre Bidanel‡ and Pascale Le Roy‡

**INRA, Station d'Amélioration Génétique des Animaux, BP 27, 31326 Castanet-Tolosan, France, †Iowa State University, Department of Animal Science, Ames, Iowa, 50011 and ‡INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy-en-Josas, France*

Manuscript received September 13, 2000
Accepted for publication June 13, 2001

ABSTRACT

We provide a theoretical framework for quantitative trait locus (QTL) analysis of a crossed population where parental lines may be outbred and dominance as well as inbreeding are allowed for. It can be applied to any pedigree. A biallelic QTL is assumed, and the QTL allele frequencies can be different in each breed. The genetic covariance between any two individuals is expressed as a nonlinear function of the probability of up to 15 possible identity modes and of the additive and dominance effects, together with the allelic frequencies in each of the two parental breeds. The probabilities of each identity mode are obtained at the desired genome positions using a Monte Carlo Markov chain method. Unbiased estimates of the actual genetic parameters are recovered in a simulated F_2 cross and in a six-generation complex pedigree under a variety of genetic models (allele fixed or segregating in the parental populations and additive or dominance action). Results from analyzing an F_2 cross between Meishan and Large White pigs are also presented.

THERE is currently much interest in the use of molecular markers to analyze the genetic basis of quantitative or "complex" traits, and an increasing number of experimental designs and statistical methods are being developed for this purpose (*e.g.*, Liu 1998). A widely used design crosses two inbred lines. In this case the quantitative trait locus (QTL) and the markers are fixed for alternative alleles. Unfortunately, completely inbred lines are available in only a few species and are certainly not available in most domestic animals or in some plants like trees. Instead, the researcher has resorted to crosses between divergent, although outbred, lines. Thus, both the marker and the QTL may be segregating in the parental lines. Furthermore neither the number of QTL alleles nor the allelic frequencies are known. It has been shown that assuming that the QTL alleles are fixed in the parental lines when this is not the case may lead to an important loss of power (Alfonso and Haley 1998; Pérez-Enciso and Varona 2000). Under additive inheritance, a mixed-model approach has been suggested for analyzing crosses between outbred lines (Wang *et al.* 1998; Pérez-Enciso and Varona 2000). However, given the well-documented phenomenon of heterosis, *i.e.*, an evidence of dominance, the assumption of additive inheritance in these methods may be an important shortcoming.

Modeling of dominance in outbred populations with inbreeding has proved to be a difficult task, given the large number of parameters that are required. Smith and Mäki-Tanila (1990) gave recursive formulas to compute the identity coefficients in a single breed. Lo *et al.* (1995) provided a general framework to model inbreeding and dominance in crossed populations. In both studies, however, no marker information was used.

The objective of this work is to present theory to analyze data from crosses of outbred lines using marker information. This theory allows dominance and inbreeding and the use of all available pedigree information. The article is organized as follows. First, we present the theory. Second, we illustrate the approach with simulated data and real data from a pig F_2 cross. The main emphasis is on F_2 crosses, given the wide popularity of this experimental design, but we also show results concerning more complex pedigrees.

THEORY

A general explanatory model for performance records is

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{g} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is a vector containing the phenotypes, $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices, $\mathbf{b}$ is the vector of fixed effects, $\mathbf{g}$, the vector that contains the genetic values, and $\mathbf{e}$ is the residuals' vector. We do not make any assumption in (1) about the pedigree structure; $\mathbf{y}$ may contain records from purebred and/or crossed individuals. In principle, any number of breeds could be accommodated, but we restricted the theory developed below to a two-breed

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

TABLE 1

Two-breed identity modes for a single individual

n_c[a]	N[b]	C[c]	F_1[d]	F_2[d]	$E(g\|N)$[e]
0	1	—	A	A	$(2p_1 - 1)a + 2p_1(1 - p_1)d$
0	2	—	A	B	$(p_1 + p_2 - 1)a + (p_1 + p_2 - 2\,p_1p_2)d$
0	2′	—	B	A	$(p_1 + p_2 - 1)a + (p_1 + p_2 - 2\,p_1p_2\,)d$
0	3	—	B	B	$(2p_2 - 1)a + 2p_2(1 - p_2)d$
1	4	A	—	—	$(2p_1 - 1)a$
1	5	B	—	—	$(2p_2 - 1)a$

[a] Number of allele pairs identical by descent.
[b] Identity mode code; a prime indicates that both modes are equivalent.
[c] Breed origin (A or B) of the allele pair if they are identical by descent.
[d] Breed origin (A or B) of each of the nonidentical-by-descent alleles.
[e] Expectation of genetic value conditional on individual identity mode N.

pedigree. The multivariate normal distribution is a very robust assumption. Thus,

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{g}\\ \mathbf{e}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{Xb} + \mu_g\\ \mu_g\\ \varnothing\end{pmatrix}, \begin{pmatrix}\mathbf{V} & \mathbf{GZ}' & \mathbf{R}\\ \mathbf{ZG} & \mathbf{G} & \varnothing\\ \mathbf{R} & \varnothing & \mathbf{R}\end{pmatrix}\right], \qquad (2)$$

where $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, $\mathbf{R}$ is a diagonal matrix with diagonal elements equal to the residual variance, σ^2, $\mu_g = \{E(g_i)\}$ is a vector containing the expected genetic values of each individual (g_i), and $\mathbf{G} = \{\mathrm{Cov}(g_i, g_{i'})\}$ is a matrix consisting of the covariances between $\mathbf{g}$ elements. The assumption of normality is required only for obtaining estimates of optimum statistical properties via Equation 5 below, but the theory developed is valid in any case. Now we need to obtain $E(g_i)$ and $\mathrm{Cov}(g_i, g_{i'})$. First, we briefly recall the theory for analyzing crossed populations developed by Lo *et al.* (1995), which was derived assuming that no marker information was available, and then we show how to obtain μ_g and $\mathbf{G}$ conditional on marker information and propose a reparameterization for QTL analysis.

The theory developed by Lo *et al.* (1995) is based on an extension to two breeds of Malécot's (1948) kinships' coefficients. Take the two alleles of a locus from a single crossed individual. The two-breed identity mode (TIM, Lo *et al.* 1995) for a single individual can take any of five mutually exclusive values, which are listed in Table 1. If the two alleles are not identical by descent, either both are from breed A ($N = 1$), or both from breed B ($N = 3$), or each allele is from a different breed origin ($N = 2, 2'$). If the alleles are identical by descent, they originate either from A ($N = 4$) or B ($N = 5$). Now consider two individuals; we can define in similar terms a pair two-breed identity mode. A schematic representation of the identity modes required to model the covariance between two individuals (see below) is depicted in Figure 1. Two related individuals can either share only one, two, or three alleles, or the four alleles can be identical by descent. (Of course two individuals can share no allele identical by descent at all, but these terms do not contribute to the genetic covariance and are not shown). Note that identity modes are grouped in Figure 1 by the number of alleles identical by descent shared by any two individuals.

Lo *et al.* (1995) showed that the expected genetic value of the ith individual is

$$E(g_i) = \sum_{j=1}^{L}\left[\sum_{k=1}^{5} E(g_{ij}/N_{ij} = k)\Pr(N_{ij} = k)\right], \qquad (3)$$

where $E(g_{ij}|N_j = k)$ is the mean genetic effect at individual i, locus j ($j = 1, L$), given that the individual two-breed identity mode, N, at locus j is k (Table 1). The

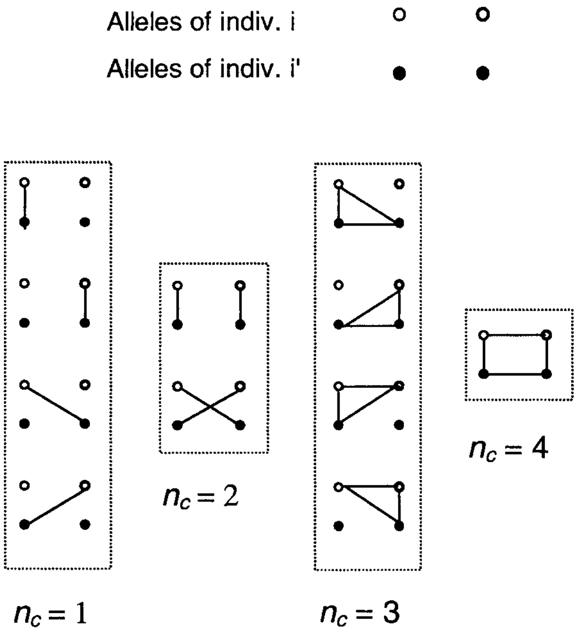

Figure 1.—Pair two-breed identity modes grouped by the number of allele pairs identical by descent (n_c). Alleles from individual i are represented by open circles and those of individual i' by solid circles. Alleles identical by descent are connected. Redrawn from Lo *et al.* (1995) but note that here only relevant identity modes for the covariance are drawn.

TABLE 2

Two-breed identity modes for two individuals

n_c[a]	M[b]	C_1[c]	C_2[c]	F_1[d]	F_2[d]	$E(g\,g'\vert N)$[e]
1	1	A	—	A	A	$[p_1^3 + (1 - p_1)^3]a^2 + [p_1(1 - p_1)]d^2 + 2p_1(1 - p_1)(2p_1 - 1)ad$
1	2	A	—	A	B	$[p_1^2p_2 + (1 - p_1)^2(1 - p_2)]a^2 + [p_1(1 - p_1)]d^2 + (2p_1 - 1)(p_1 + p_2 - 2p_1p_2)ad$
1	2′	A	—	B	A	$[p_1^2p_2 + (1 - p_1)^2(1 - p_2)]a^2 + [p_1(1 - p_1)]d^2 + (2p_1 - 1)(p_1 + p_2 - 2p_1p_2)ad$
1	3	A	—	B	B	$[p_1p_2^2 + (1 - p_1)(1 - p_2)^2]a^2 + [p_1(1 - p_2)^2 + p_2^2(1 - p_1)]d^2 + 2p_2(1 - p_2)(2p_1 - 1)ad$
1	4	B	—	A	A	$[p_2p_1^2 + (1 - p_2)(1 - p_1)^2]a^2 + [p_2(1 - p_1)^2 + p_1^2(1 - p_2)]d^2 + 2p_1(1 - p_1)(2p_2 - 1)ad$
1	5	B	—	A	B	$[p_2^2p_1 + (1 - p_2)^2(1 - p_1)]a^2 + [p_2(1 - p_2)]d^2 + (2p_2 - 1)(p_1 + p_2 - 2p_1p_2)ad$
1	5′	B	—	B	A	$[p_2^2p_1 + (1 - p_2)^2(1 - p_1)]a^2 + [p_2(1 - p_2)]d^2 + (2p_2 - 1)(p_1 + p_2 - 2p_1p_2)ad$
1	6	B	—	B	B	$[p_2^3 + (1 - p_2)^3]a^2 + [p_2(1 - p_2)]d^2 + 2p_2(1 - p_2)(2p_2 - 1)ad$
2	7	A	A	—	—	$(1 - 2p_1 + 2p_1^2)a^2 + 2p_1(1 - p_1)d^2$
2	8	A	B	—	—	$(1 - p_1 - p_2 + 2p_1p_2)a^2 + (p_1 + p_2 - 2p_1p_2)d^2$
2	8′	B	A	—	—	$(1 - p_1 - p_2 + 2p_1p_2)a^2 + (p_1 + p_2 - 2p_1p_2)d^2$
2	9	B	B	—	—	$(1 - 2p_2 + 2p_2^2)a^2 + 2p_2(1 - p_2)d^2$
3	10	A	—	A	—	$(1 - 2p_1 + 2p_1^2)a^2$
3	11	A	—	B	—	$(1 - p_1 - p_2 + 2p_1p_2)a^2 + (p_1 - p_2)ad$
3	12	B	—	A	—	$(1 - p_1 - p_2 + 2p_1p_2)a^2 + (p_2 - p_1)ad$
3	13	B	—	B	—	$(1 - 2p_2 + 2p_2^2)a^2$
4	14	A	—	—	—	a^2
4	15	B	—	—	—	a^2

[a] Number of allele pairs identical by descent. See also Figure 1.

[b] Identity mode code; a prime indicates that both modes are equivalent.

[c] Breed origin (A or B) of allele pair(s) if alleles are identical by descent.

[d] Breed origin (A or B) of each of the nonidentical-by-descent alleles.

[e] Expectation of the product of genetic values g and g' conditional on pair identity mode M.

genetic covariance between any two crossed individuals is, assuming that loci are unlinked,

$$\mathrm{Cov}(g_i, g_{i'}) = \sum_{j=1}^{L}\left[\sum_{k=1}^{25}\mathrm{Cov}(g_{ij}, g_{i'j}|M_{ii'j} = k)\Pr(M_{ii'j} = k)\right] \tag{4}$$

(Lo *et al.* 1995), where $\mathrm{Cov}(g_{i,j}, g_{i',j}|M_{ii'j} = k)$ is the genetic covariance at locus j between individuals i and i' given that their pair identity mode M is k. Lo *et al.* (1995) showed that there are 30 distinct pair identity modes when dealing with two breeds, of which 5 are zero (see Figure 4 in their article). But further, it can be shown that in fact only the first 15 identity modes in Lo *et al.* (1995) are required to model correctly **G**. Modes 16–30 in Lo *et al.* (1995) either do not contribute to the covariance or are particular cases of previously defined modes. Table 2 enumerates all 15 possible and relevant pair identity modes, classified by allele origin.

The principles of this theory can be applied to QTL detection, thus permitting the QTL analysis of populations of any pedigree structure issued from crosses between outbred populations irrespective of whether the gene action shows dominance or not. But two main obstacles persist. First, the number of genetic parameters to be estimated for *every* locus is 20, the mean plus the covariance parameters. Even if we reduced the number of parameters required in Lo *et al.* (1995), the size and pedigree structure of most QTL experiments do not suffice to obtain meaningful estimates. Second, we need to obtain the probabilities for each identity mode at every desired genome location for all pairs of related individuals, conditional on marker information.

The number of parameters to be estimated can be dramatically reduced if we assume a biallelic QTL with different frequencies in each breed. The model can now be reparameterized solely in terms of the additive (a) and dominance (d) QTL effects, plus the frequencies of each allele in breeds A and B, p_1 and p_2, respectively. The genotypic value of homozygous individuals is thus a and $-a$ for the alternative alleles, and heterozygous individuals have d as genotypic value. The conditional covariances in (4) can be obtained easily if we assume Hardy-Weinberg equilibrium in the purebred founder individuals. Consider, for instance, $M = 14$ (Table 2), *i.e.*, the case where both individuals are inbred and the locus is from breed A origin. The covariance is, dropping the subscripts in M,

$$\mathrm{Cov}(g_i, g_{i'}|M = 14) = E(g_i\, g_{i'}|M = 14) - E(g_i|M = 14)E(g_{i'}|M = 14).$$

Given that the individuals are inbred, their genotype will be a with probability p_1 (because its origin is breed A), and they share the same allele, thus

$$E(g_i\, g_{i'}|M = 14) = p_1 a^2 + (1 - p_1)(-a)^2 = a^2$$

and

$$E(g_i|M = 14) = E(g_{i'}|M = 14) = E(g_i|N = 4) = (2\,p_1 - 1)a.$$

Other conditional genetic covariances can be obtained similarly. All terms required are listed in Tables 1 and 2.

The TIM probabilities conditional on marker information can be computed via a modification of the Monte Carlo Markov chain (MCMC) approach described in Pérez-Enciso and Varona (2000) and in Pérez-Enciso *et al.* (2000b). In short, the algorithm consists of two steps, a step where unknown phases are sampled conditional on available marker information and current phases at other loci and a step where crossover positions are sampled conditional on current phases. Once crossovers are sampled it is possible to trace back the genome origins of any individual at any genome position. Thus, any identity-by-descent coefficient can be readily calculated, including the two-breed identity modes. The process just described is repeated and the mean over MCMC iterations is used to obtain Pr(M) and Pr(N) in (3) and (4). The probabilities of TIM coefficients that need to be stored are 3 (M = 7, 8, and 9) for the diagonal of a noninbred individual, 5 if it is inbred (M = 7–9, 14, 15), 9 for the off-diagonal elements if no individual is inbred (M = 1–9), and all 15 if any of the two is inbred. See Figure 1 and Table 2. We ran the MCMC for 1000 iterations. This relatively small number was good enough as the autocorrelation between samples was very small (Pérez-Enciso *et al.* 2000b). Further, we tested the algorithm with the exact analytical result in Lo *et al.* (1995) when there was no informative marker at all, and we also saw that 1000 iterations sufficed. Nonetheless, it should be borne in mind that MCMC convergence problems may exist, in particular if the percentage of missing markers is large.

Finally, it should be noted that Equation 4 assumes that loci are unlinked, which would preclude the analysis of linked QTL. However, we have shown that the covariances between loci are zero *conditional* on marker information, provided that markers are informative and distances between successive markers are small (Pérez-Enciso and Varona 2000).

We used a two-step strategy for the QTL analysis. First, the TIM coefficients were calculated at the desired genome positions. Subsequently, maximum-likelihood estimates for a, d, p_1, and p_2, plus the fixed effects, were obtained at each genome position to determine the most likely QTL location, its effect, and its frequencies. The log-likelihood is

$$L = -\tfrac{1}{2}[\text{Constant} + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{Xb} - \boldsymbol{\mu}_g)'\,\mathbf{V}^{-1}\,(\mathbf{y} - \mathbf{Xb} - \boldsymbol{\mu}_g)]. \qquad (5)$$

Note that here both $\mathbf{G}$ and $\boldsymbol{\mu}_g$ depend nonlinearly on the four parameters a, d, p_1, and p_2. In contrast, the approach in Pérez-Enciso and Varona (2000), which deals with analysis of crosses between outbred lines under an additive model, allows us to factor out each parameter separately; *i.e.*, $\mathbf{G}$ can be decomposed as

$$\sum_{i=1} \mathbf{G}_i\theta_i,$$

where θ_i are the parameters to be estimated. The nonlinearity is the price to pay for allowing dominance gene action within outbred lines. We maximized the likelihood (5) using a simplex algorithm. This algorithm is not efficient in CPU use but it is convenient because it does not require any derivatives to be calculated.

It is interesting to compare the approach followed here with other classical methods. Take $p_1 = 1$ and $p_2 = 0$. This is the model used in analyzing crosses between inbred lines. By substituting $p_1 = 1$ and $p_2 = 0$ into (3) and (4) it is straightforward to show that $\mathbf{G} = \emptyset$ and that the only terms remaining are those involved in $\boldsymbol{\mu}_g$, which are the usual regression coefficients employed in QTL analysis. If, in contrast, we set $p_1 = p_2$, we retrieve a model for analyzing outbred populations, *i.e.*, where breed origins are not taken into account. Similarly, a strict additive model can be studied by constraining d = 0. In summary, we should be able to test specific gene actions in the population under study by choosing an appropriate restriction on the parameters.

SIMULATION

Two sets of simulations, an F_2 cross and a six-generation pedigree, were simulated. The F_2 population consisted of 10 and 20 founders from each of the two breeds, 20 male and 40 female F_1 individuals, and 320 F_2 individuals. All families contributed an equal number of descendants. Two analysis options were considered: Either only performances from F_2 individuals were used (n = 320) or also records from all F_0 and F_1 individuals were available and analyzed jointly with the F_2 data (n = 410).

In addition, we also tested the method in a general pedigree. More specifically we simulated a six-discrete-generation pedigree (n = 410). It consisted of 10 and 20 founders from each of the two breeds. The individuals of the next generation were produced by mating 5 sires to two dams each, sires and dams being chosen at random with replacement (*i.e.*, an individual, male or female, could participate in more than one mating per generation), and five full-sibs per mating were generated. The exceptions were the F_1 generation, where 10 sires were chosen to produce the F_2, and the F_2, where 13 offspring per mating were generated. It was assumed that all individuals were genotyped and phenotyped. All data were included in the analysis.

The trait was assumed to be controlled by a single biallelic QTL in position 10 cM and bracketed by two markers located at 0 and 25 cM. The markers had 12 alleles, with 6 alleles specific to each breed. Hardy-Weinberg equilibrium frequencies were forced for the QTL in the founder individuals. Founder marker genotypes were sampled at random from a uniform distribution for allele frequencies. The additive genetic value was

TABLE 3

Results with the simulated F_2 cross

True parameters			Data included on generations[a]	Estimates ± SE[b]			
p_1	p_2	d		p_1	p_2	a	d
1.0	0.0	0	2	0.99 ± 0.02	0.02 ± 0.03	1.04 ± 0.10	—
			0–2	0.99 ± 0.01	0.01 ± 0.02	1.03 ± 0.08	—
1.0	0.0	1	2	0.99 ± 0.01	0.02 ± 0.03	1.05 ± 0.18	1.00 ± 0.24
			0–2	1.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.10	0.99 ± 0.14
1.0	0.5	0	2	0.99 ± 0.03	0.44 ± 0.13	0.97 ± 0.13	—
			0–2	0.99 ± 0.01	0.48 ± 0.10	1.01 ± 0.14	—
1.0	0.5	1	2	0.93 ± 0.24	0.43 ± 0.15	0.91 ± 0.60	1.06 ± 0.45
			0–2	0.95 ± 0.18	0.46 ± 0.12	0.94 ± 0.50	1.09 ± 0.32
0.5	0.5	0	2	0.56 ± 0.06	—	1.04 ± 0.11	—
			0–2	0.52 ± 0.06	—	1.02 ± 0.08	—
0.5	0.5	1	2	0.46 ± 0.10	—	0.90 ± 0.31	1.04 ± 0.34
			0–2	0.45 ± 0.12	—	0.85 ± 0.44	1.12 ± 0.29

In all cases, $a = 1$ and nongenetic variance = 1; p_1, p_2, allele frequency in the two parental breeds; a, additive effect; d, dominance deviation.

[a] 2, only F_2 data were included in the analysis; 0–2, all data analyzed jointly.

[b] Estimates obtained at the true QTL position. Average of 30 replicates.

set to $a = 1$ and d to 0 or 1. Three cases for allele frequencies were considered: $p_1 = p_2 = 0.5$; $p_1 = 1$, $p_2 = 0$; and $p_1 = 1$, $p_2 = 0.5$. All six cases considered are listed in Tables 3 and 4. The phenotype was obtained by adding a normal deviate $N(0, 1)$ to the genetic value. We report the estimates obtained by maximizing the likelihood at the true QTL position. This was done to assess the ability of the method to distinguish between alternative genetic models. The performance of the method in a chromosome scan is shown below in the real data example. Thirty replicates per case were done.

Four models were used to analyze each of the data sets generated under the six genetic situations. These were an additive model where a single allele frequency was estimated, *i.e.*, a and p ($p_1 = p_2$ forced) as parameters; second, a model containing a, d, and p; third, a model with a, p_1, and p_2 parameters; and finally a full model containing a, d, p_1, and p_2.

REAL DATA

The data were from an F_2 cross with Meishan and Large White pigs as parental populations. A comprehensive report of the experimental design and results can be found in Milan *et al.* (1998). The pedigree analyzed comprised six Large White boars, six Meishan females as founder animals, 36 F_1, and 300 F_2 individuals, which were a subset of the 1000 individuals available. Previous analysis using the approach of Haley *et al.* (1994) provided strong evidence of a QTL on chromosome 4 affecting growth, but the results with respect to backfat were less conclusive. A joint analysis of seven QTL experi-

TABLE 4

Results with the simulated six-generation pedigree

True parameters			Estimates ± SE[a]			
p_1	p_2	d	p_1	p_2	a	d
1.0	0.0	0	0.99 ± 0.04	0.01 ± 0.02	1.13 ± 0.10	—
1.0	0.0	1	0.98 ± 0.10	0.07 ± 0.18	1.03 ± 0.29	1.00 ± 0.34
1.0	0.5	0	1.00 ± 0.01	0.46 ± 0.11	1.02 ± 0.14	—
1.0	0.5	1	0.99 ± 0.01	0.45 ± 0.17	1.02 ± 0.50	1.07 ± 0.48
0.5	0.5	0	0.50 ± 0.07	—	0.99 ± 0.09	—
0.5	0.5	1	0.49 ± 0.20	—	1.04 ± 0.58	1.32 ± 0.30

In all cases, $a = 1$ and nongenetic variance = 1; p_1, p_2, allele frequency in the two parental breeds; a, additive effect; d, dominance deviation.

[a] Estimates obtained at the true QTL position. Average of 30 replicates.

ments suggested that the chromosome 4 effect on backfat in crosses involving Meishan was much smaller than in crosses with wild boar (Walling *et al.* 2000). Thus, we selected for the analysis records from backfat thickness adjusted at 80 kg and live weight adjusted for 120 days and markers from chromosome 4. Eight microsatellites were genotyped. They were located in positions 0 (S0227), 27 (SW2547), 50 (S0001), 75 (SW1089), 88 (SW270), 91 (S0214), 121 (SW445), and 141 (S0097) cM. These distances are from the average sex map. Different models were fitted at 2-cM windows between positions 50 and 90 cM. This region should largely contain the 95% confidence interval for the QTL, with maxima located in positions 75 (backfat) and 68 cM (live weight; Milan *et al.* 1998). The probability coefficients Pr(N) and Pr(M) were obtained after 1000 MCMC iterates using all marker information, even if we restricted the analysis to a specific chromosome region. The same data set was also analyzed using the regression approach in Haley *et al.* (1994), which assumes a biallelic QTL, and a within-family approach suited for a three-generation pedigree consisting of a mixture of full- and half-sib families (Le Roy *et al.* 1998). In this latter approach both the sire and the dam of the F_2 individuals are assumed to be heterozygous, not necessarily for the same alleles across families. Estimates are obtained via maximum likelihood.

RESULTS

Simulation: A first step in the analysis is to decide which is the most appropriate genetic model, *i.e.*, whether the alleles are fixed within the parental populations and whether genic action is purely additive or there is evidence of dominance. Consequently, we computed the likelihood ratio (LR) of models including dominance and/or unequal breed allele frequencies *vs.* the simplest model, *i.e.*, no dominance and equal allele frequencies in both breeds. Figure 2 shows the results corresponding to the F_2 population. The results are shown for all six parameter combinations used to generate the data. Statistics are presented for two cases, namely, whether only F_2 phenotypic records or all F_0, F_1, and F_2 records are included in the analysis. A LR test allowed us to retrieve the correct model in all instances studied. Take, for example, Figure 2a, where the null model is the true one, no LR exceeded the significance threshold. Whenever data were generated according to a purely additive model (Figure 2, a, c, and e), the LR of the model including dominance did not improve upon the additive one. Otherwise ($d = 1$), the LR clearly showed that a dominance parameter should be included in the model (Figure 2, b, d, and f). Accordingly, a LR also discriminated whether allele frequencies are equal (Figure 2, a, and b) or not (Figure 2, c–f). In cases c and d (true $p_1 = 1$, $p_2 = 0$), we also tested whether a model including parameters p_1 and p_2 improved over a model that set $p_1 = 1$ and $p_2 = 0$, with the result that the former model was not significantly better than the latter model (results not shown). Similar results, not presented to avoid repetition, were found for the six-generation pedigree.

Note, in addition, that the inclusion of parental purebred and F_1 records improves the probability of detecting the correct model as the LR of the most parsimonious correct model increases. In the particular case represented in Figure 2a ($d = 0$, $p_1 = p_2 = 0.5$), the LRs of less parsimonious models decrease when analyzing all data, giving further support to the null hypothesis model.

Average estimates of the parameters for the F_2 cross and the six-generation pedigree are in Tables 3 and 4, respectively. The estimates reported are those obtained under the correct model, the rationale being that a test has been carried out to determine which is the appropriate model, as in Figure 2. All in all, we find an excellent agreement between actual parameters and their estimates. The accuracy of allele frequency estimates was very high if alternative alleles were fixed and less so if the alleles were segregating within breeds, but still unbiased estimates were retrieved. Standard errors were, in most cases, smaller when F_0 and F_1 records were included in the F_2 pedigree analysis.

Comparing by experimental designs, the estimates of the six-generation pedigree had on average a larger standard error than those in the F_2 design when alleles were not fixed in each parental breed. This is likely to be due to genetic drift, which increases each generation, and we found a strong interrelationship between allele frequency and QTL effect estimates. In contrast, we also observed a smaller error for QTL position in the six-generation pedigree than in the F_2 design (results not presented), as expected because of a larger number of meioses in the former population (Darvasi and Soller 1995).

Pig data: The results of the comparison between alternative models on the F_2 cross pig data from Milan *et al.* (1998) are listed in Table 5. Five models were fitted. Model 1 is the model assumed in a typical regression approach including dominance; models 2 and 4 assume a pure additive action, whereas models 3 and 5 include d. Models 4 and 5 allow for different allele frequencies in each breed, whereas models 2 and 3 do not. In addition, the parameter estimates using the regression approach in Haley *et al.* (1994, model 0a) and the within-family analyses of Le Roy *et al.* (1998, model 0b) are shown as well. For backfat, the allelic action is additive, as can be seen from comparing the LR for models that include d *vs.* those that do not include d, *i.e.*, models 3 *vs.* 2 and 5 *vs.* 4. Moreover, the allele frequencies are also significantly different in each breed, as would be expected. But it is more illuminating to ask whether the breeds have alternative alleles fixed. The difference in LR between model 5 and the model where fixed alleles

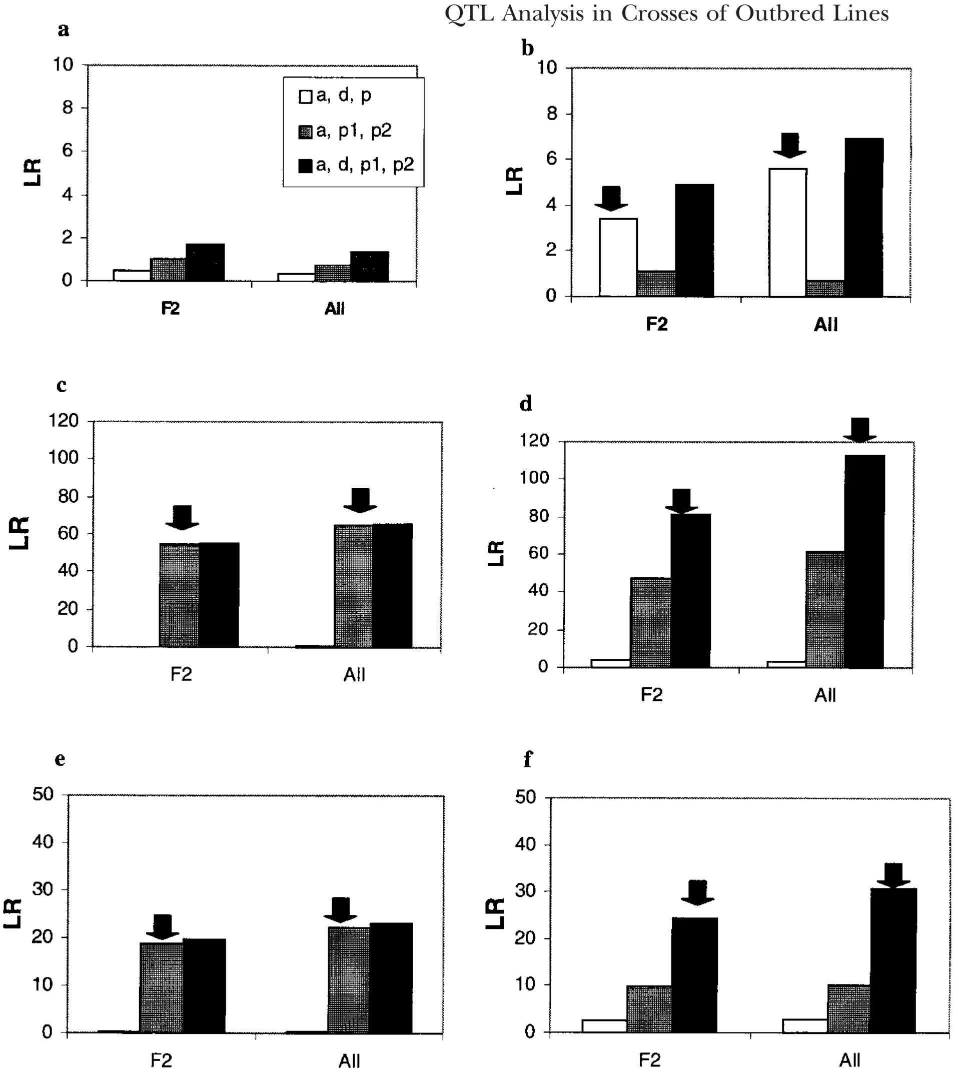

Figure 2.—Twice the log-likelihood ratio (LR) of different models with respect to the model with no dominance and equal frequencies in both breeds in an F_2 cross design. Open bar is the LR for the model with equal allele frequencies and dominance; cross-hatched bar, LR for the additive model including unequal allele frequencies; and solid bar, LR for the model allowing for unequal allele frequencies and dominance. Simulations were carried out under different genetic models: (a) $d = 0$, $p_1 = p_2 = 0.5$; (b) $d = 1$, $p_1 = p_2 = 0.5$; (c) $d = 0$, $p_1 = 1$, $p_2 = 0.5$; (d) $d = 1$, $p_1 = 1$, $p_2 = 0.5$; (e) $d = 0$, $p_1 = 1$, $p_2 = 0.5$; and (f) $d = 1$, $p_1 = 1$, $p_2 = 0.5$. Ratios were calculated at the exact QTL position. The two groups of bars above "F_2" and "All" mean only F_2 records or all F_0, F_1, and F_2 records are analyzed, respectively. The solid arrows indicate the most parsimonious correct model in each case. Results are the average of 30 replicates.

are assumed (model 1) is 6.8. The exact distribution of this ratio is not known but a chi square between 1 and 2 d.f. can be a good approximation, and even in the most conservative case (2 d.f.) it would be significant ($P < 0.05$). The analysis would thus suggest that the "fat" allele is fixed in Meishan and at low frequency but still segregating in Large White. Note that the QTL effect is underestimated in a model that forces alleles to be fixed in each breed. It is well known that power decreases in a regression approach if alleles are not fixed in each breed (Alfonso and Haley 1998; Pérez-Enciso and Varona 2000).

In contrast to backfat thickness, all statistical evidence suggests that Meishan and Large White pigs have fixed alternative alleles affecting live weight in chromosome 4. Models 4 and 5 converge to $p_1 = 1$ and $p_2 = 0$, with no increase in likelihood in model 5 with respect to model 1. It is more difficult to ascertain the effect of dominance, the difference in LR being close to significance. The regression approach provided estimates similar to those obtained under the additive model 3. Note that the QTL position estimates vary widely depending on whether dominance was included or not; QTL position changed over 10 cM according to the model chosen. Figure 3 shows a plot of LR for models that include the dominance effect or not and $p_1 = 1$ and $p_2 = 0$. It can be seen that there are two local maxima in that region, and probably the confidence interval for the QTL position comprises both maxima. In any case, this change in QTL position is particularly worrying here given that the effect of dominance borders significance. Note, in addition, that the within-family analysis agrees with the position estimated under the dominant model, whereas the between-breed regression estimate is close to that obtained with the additive model. The LRs of models that assume equal frequencies in both breeds (models 2 and 3) were nonsignificant, which contrasts to the results obtained for backfat thickness. This occurs because these models assume that there is allelic variation within breeds, which seems to be the case for backfat thickness but not for growth.

DISCUSSION

The theory developed allows us to obtain a very useful insight into QTL genic action. It allows us to diagnose

TABLE 5

Pig F_2 analysis

Trait	Model of analysis[a]	Statistics[b]	Estimates[c]				
			p_1	p_2	a	d	δ
Backfat	0*a*: regression (*H*)	6.5	—	—	−0.89	0.22	75
	0*b*: within family (LR)	38.2	—	—	−0.76	—	71
	1: $p_1 = 1$, $p_2 = 0$, a, d	18.8	—	—	−1.34	0.08	73
	2: p, a	14.7	0.66	—	−1.99	—	75
	3: p, a, d	14.7	0.58	—	−1.91	0.00	75
	4: p_1, p_2, a	24.3	0.79	0.00	−1.88	—	75
	5: p_1, p_2, a, d	25.6	0.76	0.00	−1.46	−1.02	73
Live weight	0*a*: regression (*H*)	16.2	—	—	2.91	0.18	68
	0*b*: within family (LR)	26.2	—	—	0.92	—	84
	1: $p_1 = 1$, $p_2 = 0$, a, d	17.4	—	—	2.34	2.28	84
	2: p, a	2.5	0.62	—	2.28	—	77
	3: p, a, d	4.1	0.15	—	1.94	6.44	81
	4: p_1, p_2, a	14.6	1.00	0.00	2.61	—	69
	5: p_1, p_2, a, d	17.4	1.00	0.00	2.14	1.97	81

[a] Regression (*H*) is the regression approach from Haley *et al.* (1994); within family (LR) corresponds to the within-family method described in Le Roy *et al.* (1998); other models specify the parameters, and restrictions if appropriate, that are included in the method described here.

[b] Statistics, *F*-value in the Haley *et al.* (1994) regression approach; otherwise, twice the log-likelihood ratio with respect to a model that comprises the residual variance only.

[c] Estimates obtained at the maximum-likelihood chromosome position: p_1, allele frequency in Large White; p_2, Meishan allele frequency; a, additive effect (Large White minus Meishan allele effects); d, dominance deviation; δ, QTL position in centimorgans. The a value reported in the within-family approach corresponds to the sire substitution effect averaged over the six sires.

whether the alleles are fixed within the parental populations or segregating at similar frequencies and whether the genic action is dominant or not. Unlike other indirect approaches like within-sire regression, testing can be done irrespective of the population structure, *i.e.*, number of generations, and using all available pedigree and marker information. Certainly the method can be improved; for instance, it would be desirable to use a single MCMC strategy to sample jointly the identity coefficients and the rest of the parameters. Such a strategy would provide exact estimates of the standard errors of the parameters, whereas in this likelihood framework with a simplex algorithm we need to resort to asymptotic approximations. We are currently working on a general Bayesian strategy to address this issue. Nonetheless, we have shown that the approach followed here performed quite well under a variety of genetic and pedigree scenarios (Tables 3 and 4, Figure 2).

The ascertainment of whether the QTL alleles are segregating within lines is an important issue in QTL

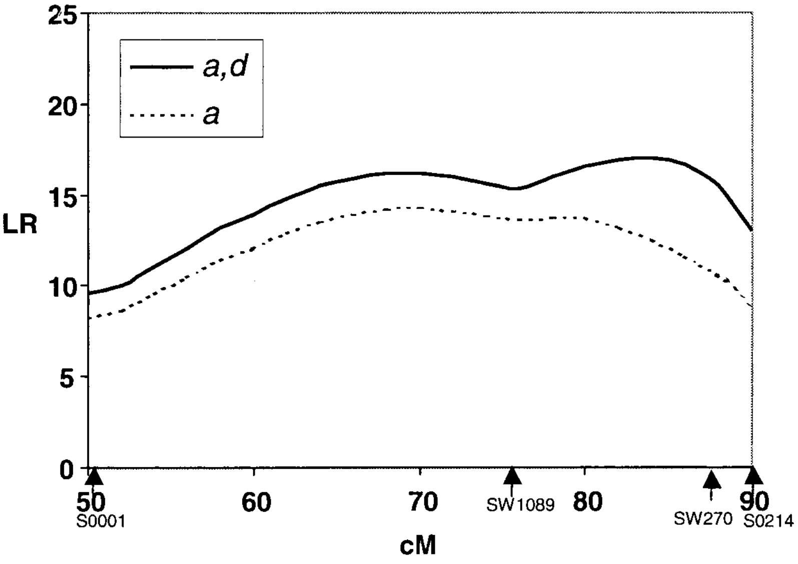

Figure 3.—Likelihood-ratio profile in chromosome 4 of live weight in the French Meishan × Large White cross. The models plotted assume alternative fixed alleles in each breed, and additive (dashed line) or additive plus dominance effects (solid line) were included. Only a fraction of the chromosome is represented; arrows indicate the positions of the microsatellites, with their names below.

identification. If a QTL is found, say, in an F_2 cross, the subsequent experimental procedure to map it finely can be very different depending on whether all F_1 are heterozygous at the QTL (alternative alleles fixed) or only a percentage are (*i.e.*, alleles segregating in the parental lines). Moreover, the presence of dominance may also alter the statistical results obtained via classical regression type methods. The power of such an approach will diminish if a recessive allele is segregating in any of the crossed populations. Certainly, the most convincing proof would be to analyze the purebreds directly, but usually the number of purebred individuals typed in experimental crosses is small and would require setting up an additional experiment. Our approach is able to extract more information than classical approaches from the already available data.

There are currently a large number of crosses between divergent lines in many different animal and plant species. Certainly, not all the parental lines utilized are completely inbred. It is thus interesting to compare the results using a regression method and the method developed here. The results presented in Table 5 represent two of the possible situations that may be encountered. For the first trait, backfat thickness, there is reasonable evidence that alleles may not be fixed in both lines. Further, the statistical analysis suggests that the QTL is segregating in Large White but not in Meishan (Table 5), which may be explained by the very small number of founder animals of the French Meishan population (Bidanel *et al.* 1989). Interestingly, it was for backfat that a joint study of seven pig F_2 crosses reported an interaction between QTL effect and experiment (Walling *et al.* 2000). One of the most parsimonious explanations for this interaction is that the QTL may be segregating at different frequencies in each breed, and the regression approach can simply not take into account this possibility. Note that model 4 in Table 5 has only one extra parameter than model 2 but that the increase in likelihood is quite important. For this trait, thus, the power will increase to a larger extent by allowing distinct allele frequencies in each breed than by including a dominance effect in the model.

In contrast to backfat, the classical model seems appropriate for live weight and there is not much to be gained by adding extra parameters to the regression model. Here some uncertainty lies on the relevance of the dominant effect, as the significance level of contrasting model 5 *vs.* 4 is $\sim P \approx 0.09$, the probability of a chi square distribution (1 d.f.) being >2.8. The evidence in favor of dominance is thus weak, in agreement with the results of the regression analysis. The QTL position estimates obtained via the within-family analyses are in agreement with those obtained via the TIM coefficients, although the average estimate is lower. In principle, one should also expect a within-family heterogeneity of substitution effects if alleles are not fixed. Nonetheless, we found that the variance of sire effect estimates using the within-family approach was similar for both traits, 0.10 and 0.08 in SD units for live weight and backfat, respectively. This is probably the result that each half-sib family is analyzed separately and thus a small number of observations is actually used to estimate each substitution effect, in contrast to the more parsimonious approach presented here where all pedigree and marker information is considered jointly.

Although beyond the scope of this article, the researcher should be aware of possible QTL position shifts according to the model of choice (Figure 3). This is pertinent especially considering that it is customary to include a dominant effect in the model in crossed-population analyses without testing for its effect. We did not carry out a joint multivariate analysis of live weight and backfat, but the fact that the allele frequencies are different for each trait would suggest that there are two linked loci. This hypothesis would be in agreement with results from Marklund *et al.* (1999) who found the QTL effect on growth, but not on fat, was diminished in a wild boar × Large White backcross when boars with different QTL genotypes for fat were progeny tested. We also applied our approach to an F_2 cross between Iberian and Landrace pigs to chromosome 4 (Pérez-Enciso *et al.* 2000a), finding evidence of alleles fixed in Iberian but not in Landrace pigs for carcass weight and fixed alleles in both breeds for backfat (M. Pérez-Enciso and A. Clop, unpublished results). All in all, this real data example illustrates the advantages of inspecting the data under different models, given that the genetic basis of all traits analyzed is not necessarily the same.

The identity modes in Lo *et al.* (1995) are a generalization of the kinship coefficients described by Malécot (1948), which are well known in quantitative genetics. They had been proved to be a very powerful instrument to model complex genetic relationships (Gillois 1964; Harris 1964; Lo *et al.* 1995) but we are not aware of any application so far in QTL studies. A definitive advantage of Malécot's coefficients is that it is straightforward to take into account other genetic situations as well. The reader can easily figure out how the relevant identity modes depicted in Figure 1 and Tables 1 and 2 could be modified to allow for, *e.g.*, imprinting or sex chromosome inheritance. We also developed a multivariate approach that allows us to distinguish between pleiotropy and linkage in a multiple QTL model; the results will be presented elsewhere. The daunting issue of obtaining the TIM probabilities conditional on marker information was solved via MCMC methods, illustrating once more the versatility of these approaches (Pérez-Enciso *et al.* 2000b). These methods are computer intensive but relatively easy to implement and program. Further, the number of parameters required was dramatically reduced by considering a biallelic locus. A biological justification of this model would be an ancestral mutation whose frequency has been changed through selection and drift at different speeds in each of the breeds studied. The pig Halothane (*Ryr1*) gene would be a classical example. But there exist as well large allelic series with a quantitative effect, like the αs_1-casein in goats (Martin *et al.* 1995). In principle, the theory can be extended to deal with a multiple breed population and more than two alleles. Each additional breed considered adds only one parameter, the allele frequency, but each additional allele increases the number by $2 + n_a$, where n_a is the previous number of alleles. This approach is not thus suitable in a large multiallelic system. A method based on analysis of variance will be more appropriate in this instance.

Finally, it should be recalled that most plant and animal individuals exploited commercially are hybrids but that their genetic evaluation is largely based on purebred performance. Thus a further application of the theory developed here, beyond the detection of QTL, will be to include molecular and performance data from hybrids in the genetic evaluation scheme. This approach can also be used to help marker-assisted introgression, where typically data from several generations are available and where dominance and inbreeding may be present.

This work began during a sabbatical visit of the senior author to Iowa State University. M.P-E. expresses his appreciation for the financial support received by Max Rothschild and Cotswold USA during this visit. The pigmap QTL project was funded by the European Union (Bridge and Biotech+ programs), Institut National de la Recherche Agronomique (Department of Animal Genetics, AIP "structure des génomes animaux" and the "Groupement de Recherches et d'Études sur les Génomes").

LITERATURE CITED

Alfonso, L., and C. S. Haley, 1998 Power of different F2 schemes for QTL detection in livestock. Anim. Prod. **66:** 1–8.

Bidanel, J. P., J. C. Caritez and C. Legault, 1989 Estimation of crossbreeding parameters between Large White and Meishan porcine breeds. I. Reproductive performance. Genet. Sel. Evol. **21:** 507–526.

Darvasi, A., and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141:** 1199–1207.

Gillois, M., 1964 La relation d'identité en génétique. Ann. Inst. Henri Poincaré **B2:** 1–94.

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.

Harris, D. L., 1964 Genotype covariances between inbred relatives. Genetics **50:** 1319–1348.

Le Roy, P., J. M. Elsen, D. Boichard, B. Mangin, J. P. Bidanel *et al.*, 1998 An algorithm for QTL detection in mixture of full and half sib families. World Cong. Genet. Appl. Livest. Prod. **26:** 257–260.

Liu, B. H., 1998 *Statistical Genomics.* CRC Press, Boca Raton, FL.

Lo, L. L., R. L. Fernando, R. J. C. 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.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité.* Masson et cie., Paris.

Marklund, L., P. E. Nyström, S. Stern, L. Andersson-Eklund and L. Andersson, 1999 Confirmed quantitative trait loci for fatness and growth on pig chromosome 4. Heredity **82:** 134–141.

Martin, P., C. Leroux, Y. Amigues, M. Jansà Pérez, F. Remeuf *et al.*, 1995 Molecular diversity of the goat alpha-S1-casein gene: impact on casein content and cheesemaking properties. Bull. Int. Dairy Fed. **304:** 12–13.

Milan, D., J. P. Bidanel, P. Le Roy, C. Chevalet, N. Woloszyn *et al.*, 1998 Current status of QTL detection in large white × Meishan crosses in France. World Cong. Genet. Appl. Livest. Prod. **26:** 414–417.

Pérez-Enciso, M., and L. Varona, 2000 Quantitative trait loci mapping in F_2 crosses between outbred lines. Genetics **155:** 391–405.

Pérez-Enciso, M., A. Clop, J. L. Noguera, C. Óvilo, A. Coll *et al.*, 2000a A QTL on pig chromosome 4 affects fatty acid metabolism: evidence from an Iberian by Landrace intercross. J. Anim. Sci. **78:** 2525–2531.

Pérez-Enciso, M., L. Varona and M. F. Rothschild, 2000b Computation of identity by descent probabilities conditional on DNA markers via a Monte Carlo Markov chain method. Genet. Sel. Evol. **32:** 467–482.

Smith, S. P., and A. Mäki-Tanila, 1990 Genotypic covariance matrices and their inverses for models allowing dominance and inbreeding. Genet. Sel. Evol. **22:** 65–91.

Walling, G. A., P. M. Visscher, L. Andersson, M. F. Rothschild, L. Wang *et al.*, 2000 Combined analyses of data from quantitative trait loci mapping studies: chromosome 4 effects on porcine growth and fatness. Genetics **155:** 1369–1378.

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

Communicating editor: C. Haley