Genet. Res. Camb. (1999), **73**, *pp.* 75–83. *With* 3 *figures.* Printed in the United Kingdom © 1999 Cambridge University Press

Mapping quantitative trait loci in a selectively genotyped outbred population using a mixture model approach

DAVID L. JOHNSON[1]*, RITSERT C. JANSEN[2] AND JOHAN A. M. VAN ARENDONK[3]

[1] *Livestock Improvement Corporation, Private Bag 3016, Hamilton, New Zealand*
[2] *Centre for Biometry Wageningen, Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Wageningen NL-6700 AA, The Netherlands*
[3] *Animal Breeding and Genetics Group, Wageningen Institute of Animal Sciences, Wageningen NL-6700AH, The Netherlands*

(*Received 26 May 1998 and in revised form 11 August 1998*)

Summary

A mixture model approach is employed for the mapping of quantitative trait loci (QTL) for the situation where individuals, in an outbred population, are selectively genotyped. Maximum likelihood estimation of model parameters is obtained from an Expectation-Maximization (EM) algorithm facilitated by Monte Carlo sampling using a Gibbs sampler. All individuals with phenotypes, whether genotyped or not, are included in the analysis where both putative QTLs and missing marker genotypes are sampled conditional on known marker information and phenotype. A simulation of a half-sib family structure demonstrates that this mixture model approach will yield unbiased estimates of the allelic effects of a QTL affecting the trait on which selective genotyping is based. Unbiased estimates were also obtained for the QTL effect on a correlated trait provided both traits were analysed jointly in a bivariate model. The procedure is also compared with a standard linear model approach. The application of these methods is demonstrated for bovine chromosome *six*, the data arising from two Holstein–Friesian families selectively genotyped for protein yield in a daughter design.

1. Introduction

DNA markers are now widely used for the detection and mapping of quantitative trait loci (QTL). Selective genotyping, the marker assay of only individuals with the more extreme phenotypes for a quantitative trait, can provide considerable savings in genotyping costs while retaining most of the statistical power for detection of QTLs affecting the trait on which the selection is based (Lebowitz, *et al.*, 1987; Lander & Botstein, 1989). For single-trait studies it will almost never be useful to genotype more than 50% of the population (the high and low tail). However, linear model estimates of the QTL effect, for which individuals without genotypic information are excluded from the analysis, will be biased by the selective genotyping (Darvasi & Soller, 1992).

A mixture model approach for the mapping of QTLs in outbred populations was presented by Jansen *et al.* (1998). This method can be applied to situations in which information about the genotype of an individual is incomplete. Incomplete information might be caused by the impossibility to trace the inheritance of an allele at a locus in an individual, unknown linkage phases between loci, unknown QTL genotype and unknown genotypes for markers. The method can, therefore, be applied to selectively genotyped data but no information is available on the properties of the estimates.

One can define a model to describe the relationship between phenotype and 'known' genotype. Since the genotype is in reality unknown, the possible genotype configurations that arise from this uncertainty then become the components of a mixture and this can be handled by an Expectation–Maximization (EM) algorithm that yields maximum likelihood estimates of the model parameters (Jansen *et al.*, 1998). A simulation study is used to investigate the performance of this mixture model approach when selective genotyping is employed within a half-sib family structure. This approach is also compared with the multi-marker regression method (Knott *et al.*, 1994).

* Corresponding author.

2. Model

For a population of N individuals let $\mathbf{y}$ denote the $N \times 1$ vector of trait values and let $\mathbf{g}$ denote the $N \times 1$ vector of genotypes where each element of $\mathbf{g}$ represents the complete genotype of marker loci and putative QTLs for an individual. Given a complete genotype $\mathbf{g}$ the conditional distribution $f(\mathbf{y} \mid \mathbf{g})$ is assumed to be multivariate normal with mean $\mu(\mathbf{y} \mid \mathbf{g})$ and variance $\mathrm{v}(\mathbf{y} \mid \mathbf{g})$. The mean $\mu(\mathbf{y} \mid \mathbf{g})$ can be modelled in terms of additivity and dominance of QTL effects. For example, consider a population comprised of a half-sib family structure, where the data have been collected on the progeny of a number of unrelated sires and y_{ij} is the trait value for the jth progeny from the ith sire. Then assuming an additive model and given the 'known' genotype

$$y_{ij} = \mu + s_i + a_{i1} q_{ij1} + a_{i2} q_{ij2} + e_{ij}, \tag{1}$$

where s_i is the (polygenic) fixed effect of the ith sire, a_{i1} is the fixed effect of the QTL allele at the first homologue of the ith sire and $q_{ij1} = 1$ or 0 depending on whether or not the jth progeny has inherited the allele of the sire's first homologue with a_{i2} and q_{ij2} defined analogously for homologue 2. The e_{ij} is a normally distributed random residual that could reflect heterogeneity of variance across families and the variable amount of information included in the trait values – for example, the number of daughters for each progeny tested sire in a granddaughter design (Weller *et al*., 1990). In this model an allelic contrast is fitted for each family. The model corresponds to a mixed inheritance model containing a polygenic effect and a multiallelic QTL (Hoeschele *et al*., 1997).

Let θ denote the vector of all parameters in the model, that is, QTL allelic effects and allele frequencies. We also denote by $\boldsymbol{h}$ the $N \times 1$ vector of observed marker data for each individual. Then the simultaneous likelihood $\mathscr{L}(\theta)$ of all the observed trait and marker data is a mixture likelihood with the possible genotypes as components

$$\mathscr{L}(\theta) = f(\mathbf{y}, \mathbf{h}) = \sum_{\mathbf{g}} P(\mathbf{g}) f(\mathbf{y}, \mathbf{h} \mid \mathbf{g}), \tag{2}$$

where $P(\boldsymbol{g})$ is the probability of a particular genetic configuration which is based on the observed marker information and is a function of recombination and allele frequencies. We note that $f(\mathbf{y}, \mathbf{h} \mid \mathbf{g}) = f(\mathbf{y} \mid \mathbf{g})$ if $\mathbf{h}$ is consistent with $\mathbf{g}$ and $f(\mathbf{y}, \mathbf{h} \mid \mathbf{g}) = 0$ otherwise. The likelihood equations, using Bayes theorem, then yield (Jansen, 1992; Jansen *et al*., 1998).

$$0 = \frac{\partial}{\partial \theta} \ln \mathscr{L}(\theta) = \sum_{\mathbf{g}} P(\mathbf{g} \mid \mathbf{y}, \mathbf{h}) \frac{\partial}{\partial \theta} \ln P(\mathbf{g}) + \sum_{\mathbf{g}} P(\mathbf{g} \mid \mathbf{y}, \mathbf{h}) \frac{\partial}{\partial \theta} \ln f(\mathbf{y} \mid \mathbf{g}), \tag{3}$$

where summation is over all possible genotypes $\mathbf{g}$ consistent with $\mathbf{h}$. As described in Jansen (1992) the first term in (3) represents genetic linkage between loci and the second term represents the phenotype–complete genotype relation.

The likelihood equations (3) can be solved by the EM algorithm. In the E-step the conditional probability $P(\mathbf{g} \mid \mathbf{y}, \mathbf{h})$ is evaluated for all possible genotypes $\mathbf{g}$ given the current parameter estimates and the observed marker information $\mathbf{h}$. The M-step involves solving each of the likelihood equations represented by the two terms in (3) using the weights $P(\mathbf{g} \mid \mathbf{y}, \mathbf{h})$ in a standard weighted regression.

A practical method for solving the likelihood equations (3) is to use a number (M) of Monte Carlo realizations:

$$\frac{\partial}{\partial \theta} \ln \mathscr{L}(\theta) = \frac{1}{M} \sum_{j} \frac{\partial}{\partial \theta} \ln P(\mathbf{g}^{(j)}) + \frac{1}{M} \sum_{j} \frac{\partial}{\partial \theta} \ln f(\mathbf{y} \mid \mathbf{g}^{(j)}), \tag{4}$$

where in the jth Monte Carlo sample, complete genotypes $\mathbf{g}^{(j)}$ are generated using the conditional distribution $P(\mathbf{g}^{(j)} \mid \mathbf{y}, \mathbf{h})$. Because of the large number of genotypic states in a population with many loci the sampling process can be facilitated by use of the Gibbs sampler (Guo & Thompson, 1992; Janss *et al*., 1995; Jansen, 1996). For the half-sib design considered here the offspring genotype can be updated in a stepwise manner for each locus and each individual. The procedure used in this paper for sampling $\mathbf{g}^{(j)}$ is outlined by Jansen *et al*. (1998).

3. Materials and methods

Data were simulated for a daughter design with 10 replicates of five sire groups with 500 daughters per group: a total of 50 sires and 25000 daughters. For each individual a 100 cM chromosome with six markers spaced at 20 cM intervals was simulated. All sires were heterozygous at all markers. A primary phenotype was simulated with a heritability of 0·3 and unit standard deviation, thus excluding variation explained by the QTL. A QTL locus with two alleles at equal frequency and positioned equidistant from markers two and three was superimposed on the polygenic background. The relative effect of the major QTL allele, $\mathrm{a}_1 = 0{\cdot}2$, was the same for each family and sires were not necessarily heterozygous at the QTL locus. A phenotype for a secondary trait was also simulated in order to study the consequences of selective genotyping based on the primary trait. The phenotypic and genetic correlations between the primary and secondary traits were both set equal to 0·7. The effect of the QTL on the secondary trait, a_2, was considered at two values, 0·2 and zero.

For genotyping, four different scenarios were

considered where either all daughters were genotyped or daughters were selectively genotyped within family:

(i) All daughters genotyped (100%).
(ii) Daughters that were extreme for the primary trait were selectively genotyped – 20% from the bottom of the distribution and 20% from the top (20%/20%).
(iii) Forty per cent of daughters were selectively genotyped but with unequal selection – 15% from the bottom and 25% from the top (15%/25%).
(iv) Truncation selection was carried out, with 80% of daughters selected from the top of the distribution only (0/80%).

For the case of selective genotyping, the marker genotypes for the unselected animals were treated as missing. Each simulated data set was analysed for different levels of genotyping. Two methods were considered for analysis of the data: the mixture model approach using the Monte Carlo EM (MCEM) method described above and the multi-marker regression approach of Knott *et al.* (1994) (REG). For the MCEM method a further comparison was made: the secondary trait, as well as being analysed as a single trait, was also included in a joint analysis with the primary trait, which we will refer to as the multiple trait analysis. In the latter analysis the sampling of genotypes was conditioned on a bivariate normal distribution of the two phenotypes and estimation of the allelic effects for each trait was based on a multiple trait linear model. In this case the multivariate normal distribution $f(\mathbf{y}|\mathbf{g})$ was expanded to include both traits as well as the linear model (1), the parameters of which are estimated using generalized least squares incorporating the environmental variance–covariance matrix between traits.

For the MCEM method all daughter phenotypes were included in the analysis with the missing marker information for unselected daughters being sampled. For the REG method only those daughters with marker data were included in the regression analysis. For comparison of the likelihood profiles between methods, the *F*-statistic of the regression method was multiplied by the degrees of freedom for the test (number of sire groups) to convert into a likelihood ratio value. The likelihood ratio values were summed over the 10 replicates. The conversion of the *F*-statistic to a likelihood ratio is not exact and therefore a comparison of likelihood profiles between methods is limited by this approximation so that a small shift in profiles does not necessarily imply that one method is more powerful than the other.

The model fitted for the MCEM method is the mixture model for a multiallelic QTL as described by (1): see model II in Jansen *et al.* (1998). Parameters include (known) recombination frequencies between markers and (unknown) marker allele frequencies within family, the mixture distribution being obtained by summing over possible genotypes. Equation (1) can also be used to describe the linear model for the REG method but with the interpretation that q_{ij1} is the probability that the *j*th progeny of the *i*th sire has received the QTL allele from the first homologue of the *i*th sire and similarly for q_{ij2}. These probabilities are calculated once on the basis of marker data only, whereas in the MCEM method the genotype (mixing) probabilities $P(\mathbf{g})$ are updated on the basis of linkage phases, marker allele frequencies, phenotype and marker observations. Further, in contrast to the MCEM method, which takes all possible linkage phases into account, the REG method determines the most likely linkage phase for each sire and in the case of equally likely phases chooses one at random. This, however, did not occur in our simulation due to the large family size. We assume residual variance to be homogeneous across families for both methods. For parameter estimation in the MCEM method, 500 Gibbs cycles were performed per EM iteration with the genotypic state at every twentieth cycle used as a Monte Carlo realization. For evaluation of the likelihood at the final iteration, 2000 Gibbs cycles were used at each of three steps for intermediate models spanning the range between the model with QTL and that without a QTL similar to that described in Jansen *et al.* (1998).

The regression of estimated QTL effect on sire QTL genotype was calculated for all scenarios. The dependent variable in the regression represented a total of 50 estimates of the QTL effect from 10 replicates with five families per replicate. The independent variable was coded as zero for homozygous sires and $+1$ or -1 for heterozygous sires depending on the phase. The expected slope of this regression is the true QTL effect, which can be compared with the simulated values. The regression was constrained through the origin. Darvasi & Soller (1992) have shown that with selective genotyping, the QTL effect as estimated from the selected population is magnified by a factor of $1+ix$ over the same effect estimated when the entire population is genotyped. Here, i is the standardized selection differential and x the standard normal deviate for the truncation point corresponding to the selected upper tail. Therefore, for a method such as REG, which does not use phenotypes of unselected individuals, one would expect the regression coefficient to be increased by a factor of $1+ix$ when comparing (equal tail) selection with no selection. For the case of (single tail) truncation selection the QTL effect is reduced by a factor of $1-i(i-x)$, the same factor as for the reduction in variance (Bulmer, 1971). This reduction is due to the increasing selection intensity with decreasing performance of different genotypes.

The MCEM and REG methods were applied to the

Table 1. *Number of daughters with phenotypes and genotypes for chromosome* six

		Marker information				
		Lower tail		Upper tail		Total
Family	Phenotypes No.	No. (%)[a]	Sel. diff.[b]	No. (%)[a]	Sel. diff.[b]	No. (%)
1	914	151 (16·5%)	−1·26	156 (17·1%)	1·27	307 (33·6%)
2	1018	133 (13·1%)	−1·24	166 (16·3%)	1·36	299 (29·4%)

[a] Number (percentage) of daughters genotyped within tails of protein yield distribution.
[b] Selection differential (average phenotype of daughters genotyped compared with average phenotype of all daughters) in standardized units.

Table 2. *Genetic markers used for chromosome* six

	Marker						
Family	BM1329	BM143	TGLA37	BM4528	BM4621	BM415	BM4311
1	1	1	1	1	1	1	1
2	1	1	0	1	0	0	1
Map, cM	0	17	23	36	40	44	57

Markers for which a sire is heterozygous are indicated by 1, otherwise 0.

investigation of marker–QTL associations in data from two Holstein–Friesian families involved in a daughter experimental design. Details of the design are presented in Table 1. The two New Zealand families comprise 914 and 1018 daughters born in 1991. Of these daughters, 307 (34%) and 299 (29%) respectively were selectively genotyped based on extreme values for protein yield within family. The actual phenotype used for selection and data analysis was the protein yield deviation, that is, protein yield adjusted for contemporary group, other fixed effects and the permanent environmental effect (VanRaden & Wiggans, 1991). The yield deviation, being an average over lactations, was further adjusted to take account of the number of lactations for each daughter to avoid, for example, overrepresentation of the more variable single lactation yields in the tails of the protein distribution. We present results for protein yield only. The seven markers, at each of which at least one of the sires was informative, and map distances are detailed in Table 2.

4. Results

For the simulated data, Fig. 1 shows the likelihood profiles for the primary trait for the two methods and four selection scenarios. The curves generally peak at the position of the QTL but also have a slightly higher likelihood value at the left end of the chromosome. The profiles for the two methods are similar for no selection and equal tail selection but tend to drift apart when selecting an unequal proportion from the two tails. There is a marked drop in likelihood for truncation selection, which is in agreement with an earlier study (Mackinnon & Georges, 1992).

Fig. 2 shows the likelihood profiles for the secondary trait for the case of no selection and equal tail selection based on the primary trait. With an effect of the QTL ($a_2 = 0{\cdot}2$) for the secondary trait the likelihood profiles were similar to those of the primary trait but at a lower absolute level. The profiles in the case of no effect of the QTL ($a_2 = 0$) were essentially flat but with those corresponding to selection sitting above those for no selection. This indicates some residual ghost QTL effect on the secondary trait due to selection on the primary trait. However, for the MCEM multiple trait analysis, the profiles in the case of selection were lower than for the corresponding univariate analysis and in particular when $a_2 = 0$ the profile was similar to that for no selection. This suggests that, for analysis of the secondary trait, it is necessary to use information on the primary trait in order to eliminate the bias generated by selection on the primary trait.

Estimates of the true QTL effect obtained from the regression of estimated family effect on sire QTL genotype are presented in Table 3. Estimates of the QTL effect for each family, 50 in total, were obtained at the position of the QTL. For the primary trait and

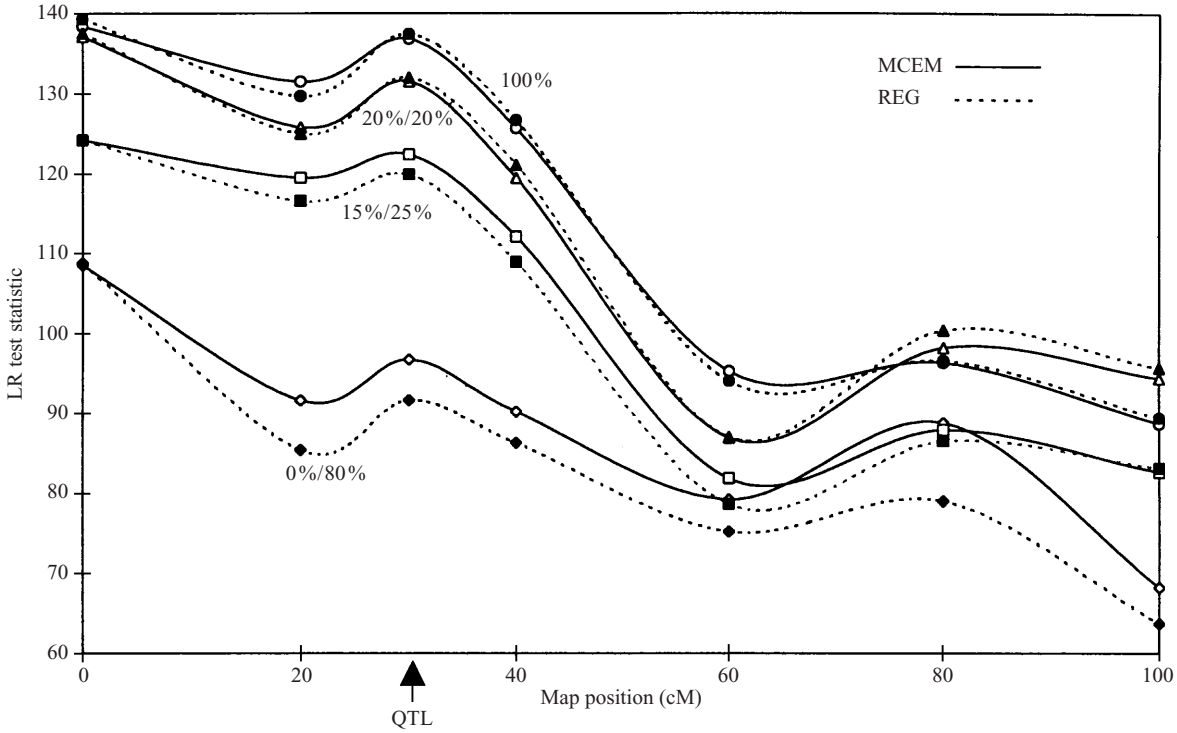

Fig. 1. QTL likelihood for the primary trait on which the following selection has been imposed: 100%, no selection; 20%/20%, top 20% and bottom 20% from the distribution; 15%/25%, 15% from bottom and 25% from top; 0/80%, 80% from top. Continuous line is the profile for the MCEM method and the dotted line for the REG method. Arrow indicates position of QTL.

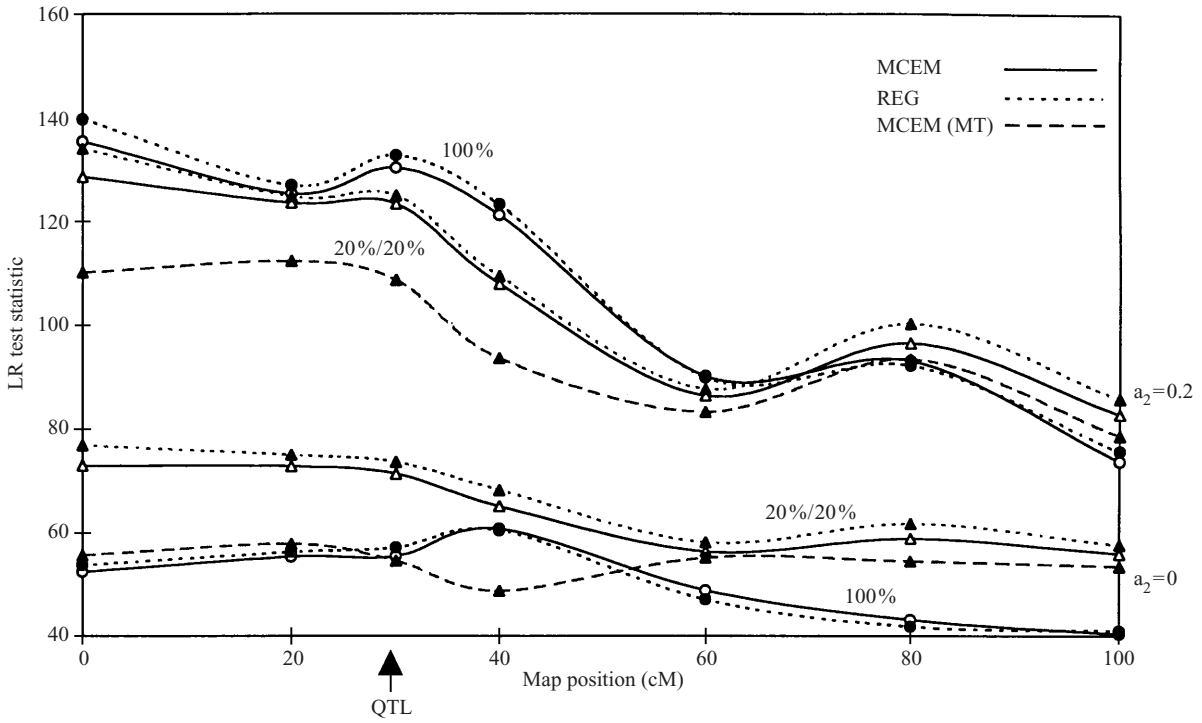

Fig. 2. QTL likelihood of secondary trait for two selection scenarios imposed on the primary trait (100% and 20%/20%) and two values of the effect of the QTL on the secondary trait ($a_2 = 0{\cdot}2$ and 0). Continuous line is the profile for the MCEM method, the dotted line for the REG method and the dashed line for the multiple trait MCEM. The plotting symbol for 100% is a circle and for 20%/20% a triangle. The multiple trait (MT) analysis is shown for the 20%/20% selection only as the profile for the 100% selection is indistinguishable from its univariate counterpart. Arrow indicates position of QTL.

Table 3. *Regression of estimated family effect on sire QTL genotype*

Trait	QTL effect	Selection	Estimated QTL effect (±SE)[a] REG	MCEM
Primary	$a_1 = 0{\cdot}2$	100%	0·169±0·023	0·166±0·024
		20%/20%	0·373±0·055	0·175±0·025
		15%/25%	0·357±0·051	0·170±0·024
		0/80%	0·092±0·020	0·146±0·031
Secondary	$a_2 = 0{\cdot}2$	100%	0·175±0·022	0·173±0·021
		20%/20%	0·327±0·043	0·204±0·026
		100% (MT)[b]	–	0·174±0·021
		20%/20% (MT)[b]	–	0·182±0·026
	$a_2 = 0$	100%	0·014±0·022	0·019±0·021
		20%/20%	0·166±0·043	0·101±0·024
		100% (MT)[b]	–	0·013±0·021
		20%/20% (MT)[b]	–	0·029±0·026

[a] Estimate (±standard error).
[b] Multiple trait analysis.

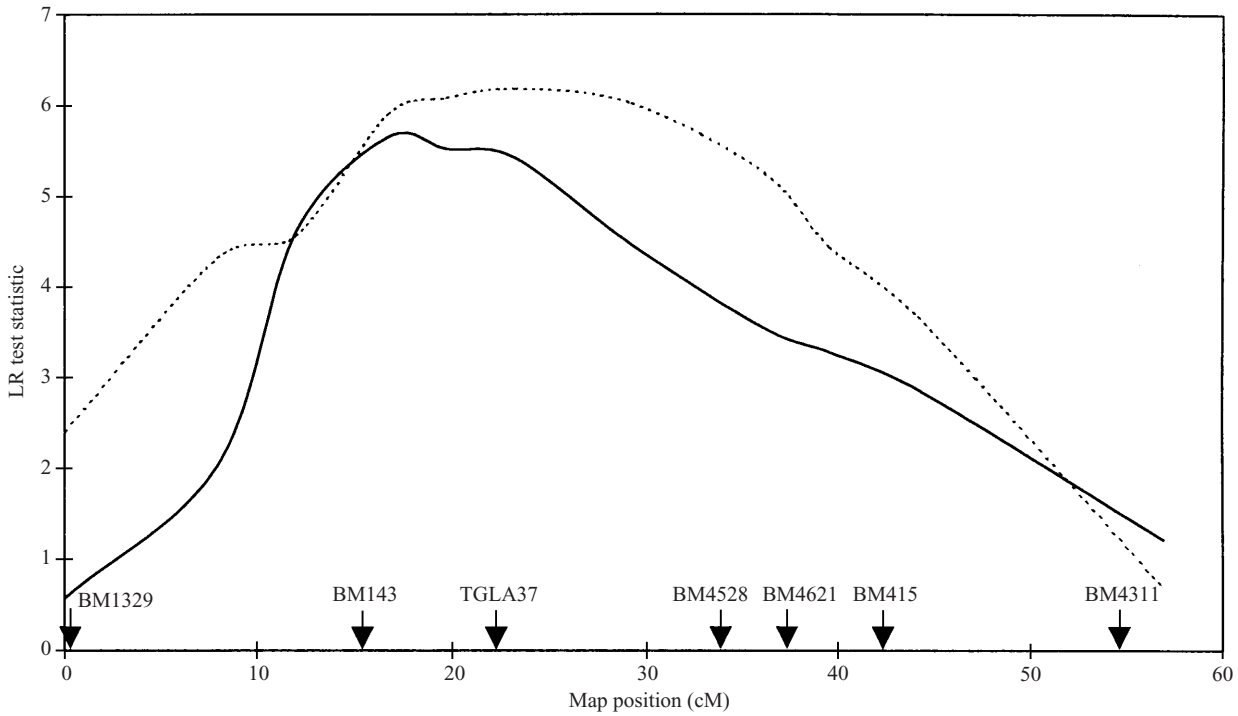

Fig. 3. QTL likelihood for protein yield from chromosome *six* in the dairy cattle experiment. Continuous line is profile for the MCEM method and dotted line for the REG method. Arrows indicate position of markers.

no selection, the estimate of the QTL from both the REG and MCEM methods was 0·17, which is not significantly different from the simulated QTL effect of 0·2. The estimated QTL effect is consistent across different selection scenarios for the MCEM method except that it is lower for the 0/80% selection (0·15) but with a higher standard error. For the REG method and equal tail selection compared with the same method for no selection, the estimate of the QTL effect is magnified by a factor of $0{\cdot}373/0{\cdot}169 = 2{\cdot}21$, which is close to the expected value of $1 + ix = 2{\cdot}17$ where the standardized selection differential (i) is 1·4, for 20% single tail selection, and the corresponding truncation point (x) is 0·84 (Darvasi & Soller, 1992). The results for 15%/25% selection are similar to those for 20%/20% selection. For the REG method and 0/80% selection the estimate of the QTL effect is reduced by the factor $0{\cdot}092/0{\cdot}169 = 0{\cdot}54$, and again this is close to the expected value of $1 - i(i - x) = 0{\cdot}58$, where, for 80% truncation selection, $i = 0{\cdot}35$ and

Table 4. *QTL effect for protein yield* (*kg*) *at position of marker TGLA37 for chromosome* six

	Method of analysis		
Family	REG	REG (adjusted)[a]	MCEM (95% CI)[c]
1	$0{\cdot}58 \pm 2{\cdot}07$[b]	$0{\cdot}22 \pm 0{\cdot}80$[b]	$-0{\cdot}43$ ($-1{\cdot}95, 0{\cdot}72$)
2	$5{\cdot}43 \pm 2{\cdot}20$[b]	$2{\cdot}09 \pm 0{\cdot}85$[b]	2·45 (1·04, 3·71)

[a] Estimates adjusted for bias due to 30% selection using scale factor of 2·6.
[b] Estimate ± standard error.
[c] Estimate and 95% confidence interval based on 1000 stochastic EM cycles.

$x = -0{\cdot}84$. This is also consistent with results of Mackinnon & Georges (1992).

For the secondary trait and no selection the REG and MCEM methods yield similar estimates for the QTL effect for both $a_2 = 0{\cdot}2$ and $a_2 = 0$ (Table 3). For the REG method, selection on the primary trait again magnifies the estimate of the QTL effect for the secondary trait. The difference between using selection and no selection in the estimates of the QTL effect for the secondary trait (using REG) is 0·152 for both $a_2 = 0{\cdot}2$ and $a_2 = 0$, and expressing this difference as a ratio of the estimated QTL effect for the primary trait under no selection (0·169) one obtains 0·9. The theoretical expectation of this ratio is $rix = 0{\cdot}82$ (Bovenhuis & Spelman, 1998), where $r = 0{\cdot}7$ is the genetic and phenotypic correlation between traits ($i = 1{\cdot}4, x = 0{\cdot}84$). This quantifies the ghost QTL effect induced by selection on the primary trait regardless of whether or not the QTL has an effect on the secondary trait. For the MCEM method under selection the estimated QTL effects for the secondary trait are also magnified relative to no selection but to a smaller extent compared with REG (0·204 vs 0·173 and 0·101 vs 0·019 for $a_2 = 0{\cdot}2$ and $a_2 = 0$, respectively). However, in the MCEM multiple trait analysis these estimates were similar to their counterparts under no selection, which reinforces the need for a joint analysis including data on which the selection is based. For the multiple trait analysis the estimates for the primary trait were essentially the same as those for the single trait analysis and are not shown in Table 3.

Fig. 3 presents the likelihood profiles for chromosome *six* and the protein yield data. The profiles are roughly similar for the two methods of analysis, with both curves peaking about the region of markers BM143 and TGLA37. The chromosome-wise empirical critical value for the likelihood ratio test statistic, determined by the permutation method (Churchill & Doerge, 1994), was 7·2 at the 10% significance level for two families based on 100000 shuffles of the phenotypes within family using the REG method. (The behaviour of the permutation test is not affected by selective genotyping, at least for the primary trait: R. J. Spelman & H. Bovenhuis, personal communication.) Further determination of the likelihood profiles within family shows that this peak is determined by family 2 only; the profile for family 1 was essentially flat and close to zero across the chromosome. The maximum likelihood value for family 2 was 5·7 at TGLA37 and the chromosome-wise critical values for this family were 4·5 and 5·9 at the 10% and 5% significance levels, respectively. The estimates for the QTL effect for the two methods at the position of TGLA37 are given in Table 4. To adjust for the bias due to selective genotyping, the estimates (and standard errors) from the REG method were scaled down by a factor of 2·6 corresponding to 30% selection (assuming equal representation from the lower and upper tails and truncation selection). These adjusted estimates, when compared with the MCEM estimates, indicate the QTL effect in family 2 at about 2 kg protein. Standard errors are not available from the EM algorithm; however, one can estimate the 'posterior' distribution of the QTL effect by using stochastic EM, a single Monte Carlo realization for each EM cycle (Jansen *et al.*, 1998). The 95% confidence intervals for the QTL effects estimated from MCEM using this distribution, based on 1000 EM cycles, are given in Table 4.

5. Discussion

The major cost in the detection of QTL with the aid of genetic markers is that due to DNA collection and typing. Selective genotyping can provide considerable cost savings, particularly in those populations where recording of phenotypes is done on a routine basis, with little loss in accuracy of detection of QTL. Linear model regression estimates of allelic effects, such as those obtained from a multi-marker regression method, are biased upwards when selection is used to genotype only those individuals that are extreme for the quantitative trait. This is due to the positive correlation between residual effects and the QTL effect in the pooled tails population that magnifies the allelic effect.

On the basis of our simulation work, the mixture model method would appear to yield estimates of gene substitution effects that are not biased by selective genotyping. This is the case not only for the primary trait on which selection is based but also for a correlated trait provided the latter is analysed jointly with the primary trait. Presumably critical to this result is the fact that the MCEM method, when sampling missing marker genotypes, takes into account not only known marker information but also all phenotypic observations for the trait.

The MCEM and REG methods give almost identical results when no selection is practised but differ in the estimates of parameters with selection. The fact that the likelihood profiles are similar between methods even for equal tail selection suggests that the REG method is still a useful and quick tool for screening for QTLs in this situation, in order to locate areas of interest for more detailed analysis, and that estimates of gene substitution effects can be adjusted for the effects of selection using the formula of Darvasi & Soller (1992). However, the adjustment assumes truncation selection and equal representation from the tails of the distribution. The MCEM method does not make any assumption on the type of selection and can also be used when part of the population cannot be genotyped due to lack of DNA. This was the case in a granddaughter design where semen samples were not available for some progeny-tested sons, resulting in families being excluded from data analysis (Spelman *et al*., 1996). Muranty & Goffinet (1997) present a simple method based on maximum likelihood for which approximate solutions are found by expanding the likelihood as a Taylor series about the maximum likelihood estimates obtained from the model assuming no effect of the QTL. Their method assumes that the effect of the QTL is small and that the genotype probabilities, $P(\mathbf{g})$, are known.

The simulated data were generated such that sires were heterozygous at all markers. Thus differences between methods in this study will be more to do with selective genotyping than with other effects resulting in incomplete marker information such as an uninformative marker locus for an entire family. A simplified analysis, less time-consuming than the mixture model method, may be possible for this simulated example. One can select the most likely linkage phase in the sires and then use an approximate expectation method that is computationally inexpensive (e.g. Knott *et al*., 1996; Muranty & Goffinet, 1997). However, one would expect the mixture model approach to be more powerful and efficient than approximate expectation methods in situations where markers are not fully informative, QTL effects are large or where the population structure is complex. We do not allow for differences between methods in estimating position, further supporting the idea that we want to quantify differences due to selective genotyping only. For the chromosome *six* data, three of the seven loci were uninformative for family 2 and so differences between methods, as represented in the likelihood profiles in Fig. 3, could be influenced by the ways in which the methods handle uninformative loci as well as the unequal representation from the tails in family 2 (Table 1).

There are other recent methods that can be used in the analysis of QTL mapping experiments (e.g. Knott *et al*., 1996; Thaller & Hoeschele, 1996; Satagopan *et al*., 1996; Uimari *et al*., 1996; Xu, 1996; Grignola *et al*., 1997). Our objective was not to compare all other methods. However, methods that deal appropriately with the selection and use all phenotypes to impute missing marker genotypes would be expected to yield estimates of QTL effects unbiased by selection.

A QTL for protein yield was identified in family 2 in the region of markers BM143 and TGLA37 on chromosome *six*. This is the same region where an effect was found for protein percentage in the American Holstein population (Georges *et al*., 1995) and in the Dutch Holstein–Friesian population (Spelman *et al*., 1996), using granddaughter designs. The two New Zealand sires involved in this study were also grandsires involved in the latter design where, for chromosome *six*, a significant QTL effect for protein percentage was found for family 1 but nothing was found for family 2 (W. Coppieters, personal communication). The number of sons in the granddaughter design was 47 and 39 for these two families, so one reason for the discrepancy between the two sets of results could be the relatively low power of the granddaughter design in this case.

Complete exploitation of all information can be obtained by the application of maximum likelihood techniques. The EM algorithm provides scope for dealing with the problem of missing QTL and marker information with the mixture model approach. Data augmentation by means of the Gibbs sampler facilitates sampling of possible genotypic states and with these 'known' genotypes standard linear regression routines can be applied. Although more computationally demanding, the MCEM method demonstrates promise as an appropriate tool for the analyses of data from QTL mapping experiments where only a proportion of the population has been genotyped.

We thank Wouter Coppieters, Michel Georges and their staff at the University of Liège (Belgium) for providing the genotypes on the two New Zealand families. We are grateful to Livestock Improvement and Holland Genetics for financial support and data access.

References

Bovenhuis, H. & Spelman, R. J. (1998). Selective genotyping to detect QTL for multiple traits in outbred populations. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, *Armidale*, *Australia* **26**, 241–244.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *American Naturalist* **105**, 201–211.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Darvasi, A. & Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics* **85**, 353–359.

Georges, M. D., Nielsen, D., Mackinnon, M., Mishra, R., Okimoto, R., *et al*. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**, 907–920.

Grignola, F. E., Hoeschele, I. & Tier, B. (1997). Mapping quantitative trait loci via residual maximum likelihood. 1. Methodology. *Genetics*, *Selection*, *Evolution* **28**, 479–490.

Guo, S. W. & Thompson, E. A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics* **51**, 1111–1126.

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

Jansen, R. C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252–260.

Jansen, R. C. (1996). A general Monte Carlo method for mapping quantitative trait loci. *Genetics* **142**, 305–311.

Jansen, R. C., Johnson, D. L. & Van Arendonk, J. A. M. (1998). A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics* **148**, 391–399.

Janss, L. L. G., Thompson, R. & Van Arendonk, J. A. M. (1995). Application of Gibbs sampling for inference in a major gene–polygenic inheritance model in animal populations. *Theoretical and Applied Genetics* **91**, 1137–1147.

Knott, S. A., Elsen, J. M. & Haley, C. S. (1994). Multiple marker mapping of quantitative trait loci in half sib populations. *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production*, *Guelph*, *Canada* **21**, 33–56.

Knott, S. A., Elsen, J. M. & Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**, 71–80.

Lander, E. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lebowitz, R. J., Soller, M. & Beckmann, J. S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics* **73**, 556–562.

Mackinnon, M. J. & Georges, M. A. J. (1992). The effects of selection on linkage analysis for quantitative traits. *Genetics* **132**, 1177–1185.

Muranty, H. & Goffinet, B. (1997). Selective genotyping for location and estimation of the effect of a quantitative trait locus. *Biometrics* **53**, 629–643.

Satagopan, J. M., Yandell, B. S., Newton, M. A. & Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. *Genetics* **144**, 805–816.

Spelman, R. J., Coppieters, W., Karim, L., Van Arendonk, J. A. M. & Bovenhuis, H. (1996). Quantitative trait loci analysis for five milk production traits on chromosome *six* in the Dutch Holstein–Friesian population. *Genetics* **144**, 1799–1808.

Thaller, G. & Hoeschele, I. (1996). A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci. 1. Methodology. *Theoretical and Applied Genetics* **93**, 1161–1166.

Uimari, P., Thaller, G. & Hoeschele, I. (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* **143**, 1831–1842.

VanRaden, P. M. & Wiggans, G. R. (1991). Derivation, calculation, and use of national animal model information. *Journal of Dairy Science* **74**, 2737–2746.

Weller, J. I., Kashi, Y. & Soller, M. (1990). Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science* **73**, 2525–2537.

Xu, S. (1996). Computation of the full likelihood function for estimating variance at a quantitative trait locus. *Genetics* **144**, 1951–1960.