Animal Science 1998, 67: 257-268

1357-7298/98/7215025 7$20-00

A strategy for QTL detection in half-sib populations

D. J. de Koning[1,4], P. M. Visscher[2], S. A. Knott[3] and C. S. Haley[1]

[1]*Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS*
[2]*Institute of Ecology and Resource Management, University of Edinburgh, West Mains Road, Edinburgh EH9 3JG*
[3]*Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT*
[4]*Wageningen Agricultural University, Department of Animal Breeding, PO Box 338, 6700 AH Wageningen, The Netherlands*

Abstract

A statistical analysis strategy for the detection of quantitative trait loci (QTLs) in half-sib populations is outlined. The initial exploratory analysis is a multiple regression of the trait score on a subset of markers to allow a rapid identification of possible chromosomal regions of interest. This is followed by multiple marker interval mapping with regression methods within and across families fitting one or two QTLs. Empirical thresholds are determined by experiment-wise permutation tests for different significance levels and empirical confidence intervals for the QTLs' positions are obtained by bootstrapping methods. For traits with evidence for a significant single-QTL effect, an approximate maximum likelihood analysis is performed to obtain estimates of QTL effect and the probability of the QTL genotype for each parent of a group of half-sibs. The strategy is demonstrated in an analysis of previously published data on chromosome 6 and five production traits from a granddaughter design in dairy cattle. The results confirm and extend evidence for QTLs affecting protein percentage. Informativeness of markers limited the possibility of mapping more than one QTL on the same linkage group.

Keywords: *dairy cattle, gene mapping, milk protein, quantitative trait loci.*

Introduction

Many traits of economic interest in plants and animals are of a quantitative nature. That is, the observed phenotypes are continuously distributed and reflect the action of many quantitative trait loci (QTLs) together with environmental effects. The availability of genetic markers has allowed experimental studies in a number of species of the nature and location of some of these QTLs. Such experiments are often based on crosses between inbred lines with large phenotypic differences. Methods of analysis have focused on identifying single QTLs of relatively large effect against an assumed background of no genetic variation on the linkage group. Although methods for the joint mapping of two or more QTLs have been developed and applied (e.g. Haley and Knott, 1992; Martinez and Curnow, 1992; Jansen, 1993), distinguishing the effects of multiple linked QTLs from those of a single QTL is difficult using these methods. More recently, Visscher and Haley (1996) have looked explicitly at alternative genetic models and discussed how one might test for the failure of the models with only a single QTL.

Genetic linkage maps have now been developed for the major livestock species. The use of crosses between genetically divergent lines to map QTLs, however, will often be difficult in livestock and the study of existing populations will often be the only practical option. Large half-sib families exist in livestock species where artificial insemination is extensively used and a relatively small group of elite sires has many offspring in the population. Dairy cattle provide the most striking example of this structure. 'Granddaughter' and 'daughter' designs for QTL detection, proposed by Weller *et al.* (1990), make use of this existing family structure and the fact that the phenotypes are collected on a routine basis for progeny testing. In such designs evidence for QTLs comes from segregation within paternal half-sib families. In the daughter design, such evidence comes from co-segregation of marker alleles and performance in the daughters. In the

granddaughter design, the evidence comes from co-segregation of marker alleles in the sons with performance assessed from records on large numbers of their daughters (granddaughters of the original sire).

Initial QTL studies in cattle were based on the daughter design and analysed associations with markers individually (e.g. Neimann-Sorensen and Robertson, 1961). The development of marker maps allows information from linked markers to be combined and the advantages of this for analyses of half-sib data have been demonstrated (Knott *et al.*, 1994 and 1996). Recently, granddaughter designs have been applied in several studies to detect QTLs in dairy cattle. Ron *et al.* (1994) use a single marker approach whereas Georges *et al.* (1995), Spelman *et al.* (1996) and Vilkki *et al.* (1997) used information from multiple markers.

In this paper, we present a step-wise approach to analyse data from multiple markers in half-sib populations. The approach uses exploratory analyses similar to those employed by Visscher and Haley (1996), in which trait scores are regressed onto selected marker information in an attempt to determine which chromosomal regions contribute to variation in the trait. Where warranted by the earlier analyses, these are followed by least-squares based interval analyses designed to further test and estimate the effects of major single QTLs. Finally, we use approximate maximum likelihood analyses to further refine our understanding of previously detected effects. We will demonstrate the use of the methods with the results of an analysis of data from a granddaughter design in cattle with marker data from bovine chromosome 6.

Methods

Approaches

The basic approach is developed from methods used previously for analyses of data from inbred line crosses (e.g. Haley and Knott, 1992), outbred line crosses (e.g. Andersson *et al.*, 1994; Haley *et al.*, 1994) and half-sib analyses (Knott *et al.*, 1994 and 1996). Essentially, the available marker information is used to calculate 'virtual' markers at chosen points (e.g. 1 cM (centiMorgan) intervals) in the linkage groups (essentially, the probabilities of each genotype at each point conditional on available marker information). The calculated 'virtual' markers then form the basis of further analyses performed via least squares, maximum likelihood or some other method. In this study we calculate the 'virtual' markers as outlined for half-sib families in Knott *et al.* (1996). We assume the common parent of a group of half-sibs will be male. In this case, we are only interested in using information on segregation from the sire, as in a true half-sib structure each dam has only one progeny and therefore there is no information from segregation within dams. In the method used, the marker data is first used to identify the most likely phase of the two sire gametes based upon progeny information from pairs of adjacent informative (i.e. heterozygous) markers. (NB in this, as in other interval analyses, we assume that the order and distances between markers are known in advance). Following this, the probability of each progeny inheriting each of the two sire gametes at each chosen point in a linkage group is calculated based on the sire genotype, the progeny genotype and, if available, the dam genotype. These probabilities (our 'virtual' markers) then form the basis of our further analyses. It should be noted that the amount of information in the marker genotypes can be assessed from these probabilities (Spelman *et al.*, 1996). At the position of a fully informative marker we know for sure which sire allele each progeny has inherited, so the probability of inheriting one allele will be unity and the probability of inheriting the other will be zero. As we move away from the position of a marker or if markers are less than fully informative the probability of the most likely genotype drops below unity. At points where there is no marker information (i.e. unlinked to any marker) a progeny has equal probability of being either sire genotype. The variance of these probabilities can be used to give a picture across the genome of information per sire or across sires, as shown by Kruglyak and Lander (1995).

Weighting the analysis

One of the most likely applications of a half-sib QTL analysis will be to data generated from a granddaughter design as applied to dairy cattle (Weller *et al.*, 1990). In such a design the common parent is the grandsire of the animals with phenotypic records but phenotypes of the sons may be expressed in terms of daughter yield deviations (DYD). The daughter yield deviation of a son is the unregressed weighted average of his daughters' performance, expressed as a deviation from the population mean (Van Raden and Wiggans, 1991). The variance of DYD depends on the number of daughters with records and if there are big differences in number of offspring between sons, their observations should be weighted to account for variance differences. These weights can be expressed as the reliability of the DYD estimated from the progeny testing data following Georges *et al.* (1995).

If A is the reliability of a son's breeding value (or transmitting ability), and B is the reliability of the

average breeding value of his parents (the so-called parent average), then the reliability pertaining to the daughters of the son, i.e. the reliability of the DYD, can be written as,

$$R = [A/(1-A) - B/(1-B)] \;/\; [A/(1-A) - B/(1-B) + 1]$$

which is a simplification of the equation provided in the appendix of Georges *et al.* (1995).

Exploratory regressions on multiple markers

The first analyses of a linkage group are simple and fast protocols to determine whether the chromosomal region under study is associated with variation in the recorded traits. The analyses can be done with standard statistical packages; Genstat 5 Committee (1993) was used in this study. To undertake the analyses, first, the locations of informative and evenly spaced markers are identified from those available in the data set. There are currently no hard and fast rules that dictate the number and spacing of markers to select. The number of markers to be selected depends on the available markers and their informativeness and the number of recorded animals. If too many markers are selected the analysis will take up a significant proportion of the degrees of freedom and information from closely linked markers is highly correlated. On the other hand, selection of widely spaced markers may not provide information on regions between markers and gives little opportunity to test for multiple linked QTLs. Theoretical calculations have indicated that markers which are spaced every 25 cM or so should explain most of the variation on a chromosome (e.g., Dekkers and Dentine, 1991; Visscher, 1996).

In the first analysis the data are regressed on to the marker locations selected for a linkage group. If a marker is uninformative in a particular individual, it is replaced by the 'virtual' marker probability calculated for that position based on other marker information. The model includes the effect of sire, marker genotype nested within sire and residual, with the test of marker within sire providing evidence on the presence of genetic variance associated with the region (Neimann-Sorensen and Robertson, 1961). For each marker we regress on to the probability of only one of the sire alleles, as the probabilities of the two sire alleles are completely confounded (and sum to unity). (NB if the data are from a granddaughter design, we can replace sire in the foregoing by grandsire.) The joint regression onto all marker locations in a linkage group is equivalent to the 'chromosomal test' of Visscher and Haley (1996), testing for the presence of genetic variation associated with that chromosome or linkage group.

The model for these exploratory analyses is:

$$Y_{ij} = \mu_i + \sum_{k=1}^{n} b_{ik} m_{ijk} + e_{ij}$$

where Y_{ij} is the DYD for son j of grandsire i, μ_i is the mean for the half-sib offspring of grandsire i, b_{ik} is the effect of one of the paternal alleles for marker k within family i, m_{ijk} is the probability for son j of inheriting that paternal allele of marker k conditional on the marker genotypes and e_{ij} is the residual effect for son j. Under the null hypothesis of no genetic variation of the trait associated with the linkage group under study only a family mean is fitted.

If an effect of a chromosome is found, further analyses can be used to identify whether there is one or more regions affecting the trait. The regression on all selected marker locations is compared with the regression on every pair of adjacent markers. If there is a single QTL located on the linkage group (or group of QTLs unseparated by a marker), then its two flanking markers absorb the QTL effect and regression on those two markers should not be a significantly worse fit than regression on all markers. Regression on pairs of markers at a reasonable distance from the QTL should fit significantly worse than regression on all markers. Where more than one important QTL affects the trait, there will be no single pair of adjacent markers that accounts for as much variance as do all markers jointly. Thus in a regression analysis we can look for improvement in fit comparing including all markers from a linkage group with each pair of markers in turn.

An alternative approach is dropping each pair of adjacent markers in turn and comparing regression on the reduced number to regression on all selected markers. Dropping pairs of markers that do not flank the QTL should not affect the fit of the model.

Thus, with these tests, if there is genetic variation associated with the linkage group regression on all markers should be significant. If the effects are due to a single clear-cut QTL in a linkage group, fitting the best pair of adjacent markers should be as good as fitting all markers and dropping any other pair of markers should not result in a significantly worse fitting model. If two or more regions (QTLs) of the linkage group affect the trait, dropping two or more pairs of adjacent markers should result in a worse fitting model. In practice, the power of the approach will be affected by the sample size and results may be less clear-cut.

Interval mapping

Linkage groups that gave evidence in the first analyses to account for genetic variation in any of the

traits can be further analysed applying interval mapping. Interval mapping was originally implemented via maximum likelihood (Lander and Botstein, 1989). Martinez and Curnow (1992) and Haley and Knott (1992) describe interval mapping using regression methods which are computationally less demanding than maximum likelihood methods and gave very similar results. Knott *et al.* (1994 and 1996) and Haley *et al.* (1994) extend these methods to outbred populations. In this study we apply the method described in Knott *et al.* (1996) with the extension that weighted least squares are used to account for different numbers of recorded daughters per son. At each 1 cM position, the trait scores from the sons are regressed on the virtual marker values (i.e. the probabilities of inheriting a given grandsire allele), calculated as described previously.

The model used is essentially that given previously

$$(\text{i.e. } Y_{ij} = \mu_i + \sum_{k=1}^{n} b_{ik} m_{ijk} + e_{ij}).$$

However, the summation is now over the number of QTLs included in the model (one or two in the analyses performed here) and m_{ijk} refers to the virtual marker value at the particular chromosomal position or positions being considered. The analysis is nested within families which allows a different linkage phase between marker genotype and putative QTLs in the different grandsires. F ratios are calculated across families by comparing the mean square due to the putative QTL effect with the residual mean square. In a one QTL model, the location where the F ratio has a maximum is the most likely position of a single QTL and model parameters are estimated at this position (Knott *et al.*, 1994 and 1996).

An alternative analysis testing for heterogeneity between families can be performed in which a QTL is fitted at the best position within each family, rather than the single overall best position. If the combined within family analysis explains significantly more variance in the trait than the across family analysis, it suggests different QTLs are segregating in the different families.

If there is evidence for more than one QTL affecting the trait on the linkage group, the previous model can be extended to fit two QTLs at all possible combinations of positions (Haley and Knott, 1992). It is impossible to map multiple QTLs at positions flanked by the same informative markers (Whittaker *et al.*, 1996). Since informative markers differ between families, there is no solution common to all families. Therefore, the model was reduced to fit just one QTL for a family if a dependency was detected within that family. A test to compare fitting one *versus* two QTLs is the F ratio calculated from the mean squares of the one and two QTL analyses. Dependencies in the two QTL analyses, which result in only a single QTL being fitted in some families, will tend to make the test for the presence of two QTLs more conservative.

Approximate maximum likelihood interval mapping

Where the analyses suggest a single QTL is segregating, the least-squares analyses provide estimates of the gametic effects that differ from family to family. If we are prepared to assume that a single QTL with two alleles is segregating, we can obtain better estimates of the QTL effect and posterior probabilities of the QTL genotype of each grandsire (i.e. whether heterozygous or homozygous for the QTL alleles). To do this in a granddaughter design the probabilities for each son at all locations of inheriting the first grandsire gamete can be used in an approximate maximum likelihood analysis, and similarly in a daughter design (Knott *et al.*, 1996). The model applied assumes a single QTL with only two alleles and no dominance. Furthermore, it is assumed that the QTL does not have a major effect on the distribution of phenotypes within half-sib families. The likelihood is optimized at fixed locations along the chromosome and compared with the likelihood of the null hypothesis of no QTL. The test statistic in this study is twice the log likelihood ratio of the two models. The position with the overall highest statistic is the most likely position of a putative QTL. Using the maximum likelihood estimates, the posterior probabilities of whether the QTL is homozygous or heterozygous can also be calculated for each grandsire.

The likelihood for a QTL at a given position requires three parameters: the frequency of sires homozygous at the QTL (p), the substitution effect (α) of the QTL and the residual variance (σ_w^2) within groups of progeny inheriting one grandsire gamete. The likelihood appropriate to the granddaughter design derived based on Knott *et al.* (1996) is:

$$L = \prod_{i=1}^{s} [p \times LQQ_i + \frac{1-p}{2} \times LQq_i + \frac{1-p}{2} \times LqQ_i]$$

where

$$LQQ_i = \prod_{j=1}^{n_i} f(z_{ij}; 0, \sigma_w^2 / R_{ij}),$$

$$LQq_i = \prod_{j=1}^{n_i} [m_{ij} f(z_{ij}; \alpha/2, \sigma_w^2/R_{ij}) + (1 - m_{ij}) f(z_{ij}; -\alpha/2, \sigma_w^2/R_{ij})],$$

$$LqQ_i = \prod_{j=1}^{n_i}[m_{ij} f(z_{ij};\alpha/2,\sigma_w^2/R_{ij}) + (1 - m_{ij}) f(z_{ij};\alpha/2,\sigma_w^2/R_{ij})],$$

are the homozygous and two heterozygous contributions to the likelihood, z_{ij} is the corrected DYD for sire j, son of grandsire i, m_{ij} is the conditional probability of son j inheriting the first gamete from the grandsire i at the considered position, R_{ij} is the reliability of the DYD of the son j of sire i and n_i is the number of sons of grandsire i.

$$f(z;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} \exp[-(z-\mu)^2/2\sigma^2]$$

is the density function of the normal distribution.

The posterior probabilities of the QTL genotypes of each grandsire can then be obtained using the maximum likelihood estimates. The probability that grandsire i is homozygous is:

$$p \times LQQ_i / [p \times LQQ_i + \frac{1-p}{2} \times LQq_i + \frac{1-p}{2} \times LqQ_i]$$

and the probability it is heterozygous is one minus this value.

Two ways of assigning genotypes to the grandsire are to pick the one with the highest probability or to only assign a genotype if the probability exceeds a certain value (e.g. 0·75) (Knott *et al.*, 1991).

Significance thresholds

Lander and Kruglyak (1995) suggest that, no matter how many linkage groups are analysed, a genome-wide scan should always be assumed when setting the significance threshold. This stringent threshold takes account of the large number of tests being performed. Lander and Kruglyak (1995) further distinguish between significant overall linkage with a genome-wide risk of $P < 0{\cdot}05$ for type I errors and suggestive linkage where one false positive is expected in a genome wide scan. A further option is to adjust thresholds for the analysis of multiple traits. Multiple correlated traits can be reduced to a lower number of independent traits explaining all the variance by a principal component analysis on the phenotypic values or the correlation matrix between the different traits (Chatfield and Collins, 1989).

For the exploratory analyses of single linkage groups, the 0·05 significance threshold can be obtained from standard tables. The overall significance used for the chromosomal test when undertaking a genome-wide scan of n independent linkage groups (or n [linkage groups × independent traits]) can be calculated from the nominal significance level applied to a single linkage group following Bonferoni:

$$P_{overall} = 1 - (1 - P_{nominal})^n.$$

A very good and simple approximation for the solution of this equation is,

$$P_{nominal} \approx P_{overall}/n.$$

The suggestive significance level can be obtained from the binomial distribution as:

$$P_{suggestive} = 1/n.$$

The stringent overall significance threshold is not appropriate once a linkage group associated with variation has been identified and further analyses are being performed based on pairs of adjacent markers. Here we would propose to use the nominal significance threshold of $P < 0{\cdot}05$.

The levels of overall and suggestive significance are also applied in the interval mapping analyses. Empirical significance thresholds are determined using permutation tests following Churchill and Doerge (1994) for the interval mapping analyses fitting one and two QTLs. The thresholds were determined for the across family analyses but could also be applied to obtain an empirical significance threshold for an individual family. Within families, the sons' phenotypes are permuted against their marker genotypes. It should be noted that the empirical thresholds obtained by permutation for a two QTL analysis assume a null hypothesis of no QTLs *versus* an alternative hypothesis of two QTLs affecting the trait. Therefore, these thresholds give no evidence to distinguish between one, two or more than two QTLs affecting the trait. A conservative approach might be to use the single QTL thresholds for testing the improvement of a two QTL model over a single QTL model.

The maximum likelihood analyses are relatively computationally demanding making it difficult to use permutation tests to define a significance threshold. This is not problematical in the series of analyses performed here, as maximum likelihood is used to provide estimates for QTLs detected in other analyses.

Confidence intervals for QTL positions

In has proved difficult to obtain confidence intervals on estimated positions of QTLs. Lander and Botstein (1989) suggest that the area bounded by a 1 or 2 unit drop in logarithm of odds (LOD) score provides an asymptotically 96·8 and 99·8% confidence interval. However, van Ooijen (1992) shows that such confidence intervals can often be anti-conservative

and Visscher *et al.* (1996) propose bootstrapping to obtain empirical threshold values for QTL positions as an alternative to the LOD drop-off methods and show that it produces good estimates of the confidence interval. For half-sib analyses resampling should be performed within family, so from each half-sib family with n sons n individuals are sampled with replacement. The analyses of N resampled replicates allow top and bottom percentiles to be identified to provide limits for the confidence interval.

Example

Data

The strategy was applied to a granddaughter design exploring linkage between bovine chromosome 6 and five production traits. These data were generated in the MILQTL project, a joint project of Liege University, Massey University and Wageningen Agricultural University, and were distributed among interested groups for analysis. The design consists of 20 grandsire families of the Dutch Holstein-Friesian population with nine to 139 half-sib sons. Genotypes were obtained for nine microsatellite markers on chromosome 6, spanning a 95 cM interval. Table 1 gives an overview of the marker information in the different families and the number of sons for each family.

The phenotypic data consisted of DYDs for milk yield, fat yield, protein yield, protein percentage and fat percentage. Further information that was provided dealt with the parental averages for the five traits, reliabilities for the breeding values of the sons and reliabilities for the parental averages of their parents, and number of daughters for each son. These data were used to calculate the reliability of the sons' DYDs and apply these as weights in all of the least-squares analyses.

Computation

Exploratory analyses were performed using the statistical package Genstat 5 Committee (1993). Least-squares and maximum likelihood (ML) interval mapping analyses were performed with programs written in Fortran 77. ML analyses used a grid search of QTL positions along the chromosome, with a Simplex routine used to maximize the likelihood over other parameters for the fixed QTL positions.

Results

Principal component analysis revealed, as expected, that the five traits can be reduced to three independent traits that account for 99% of the variance. For exploratory analyses of a single trait with 29 autosomes the suggestive and significant thresholds would be $P < 0{\cdot}034$ and $P < 0{\cdot}002$,

Table 1 *Information on the granddaughter design and markers*

	Marker (position in cM)									
Grandsire	1 (0)	2 (13)	3 (20)	4 (31)	5 (41)	6 (52)	7 (54)	8 (58)	9 (94)	n
1		#			#		#	#		13
2		#	#		#	#			#	12
3					#		#	#		16
4		#			#		#	#	#	31
5		#			#	#		#		42
6	#	#		#			#		#	139
7	#	#	#	#	#	#	#	#		13
8		#		#		#			#	53
9		#	#	#	#	#				23
10		#	#	#	#		#	#	#	71
11		#		#		#	#	#	#	26
12	#	#	#	#	#	#		#	#	12
13		#	#		#	#	#	#		73
14	#	#			#		#	#	#	59
15		#			#		#	#		22
16		#		#	#	#	#	#		38
17		#	#				#			15
18		#		#		#	#	#	#	14
19				#					#	16
20		#	#	#	#		#	#		9

# Denotes that a grandsire is heterozygous for a marker and n is number of half sibs sons in the family.

respectively. For analysis of three independent traits the suggestive level and significant thresholds would be $P < 0{\cdot}011$ and $P < 0{\cdot}0006$, respectively.

For the exploratory analyses we used virtual markers at the positions of markers 2, 4, 5, 7 and 9 (NB this is equivalent to using the marker genotype for individuals in which the marker is fully informative). Using these markers gave an average of 15·3 cM between markers (Table 1). Each of these markers was informative in at least 10 of the 20 grandsire families and the markers were reasonably spaced. The multiple regression of the trait score on these five markers revealed an effect on milk yield which approached the single trait suggestive level ($P = 0{\cdot}042$) and an overall significant effect for protein percentage ($P < 0{\cdot}0006$).

There was no significant evidence for an effect on fat or protein yield or on fat percentage.

Regression on all markers was then compared to regression on pairs of adjacent markers. For protein % this revealed that the model with all five marker locations included was always significantly ($P < 0{\cdot}005$ in all cases) better than fitting any pair of adjacent markers. There is thus not one clear region that accounts for the variance explained by the five markers, i.e. there must be more than one region accounting for the effect. For the other traits, no single pair of adjacent markers explained significantly less of the variance than the five markers altogether. The alternative analysis of protein % in which pairs of adjacent markers were dropped from the full model indicated that each pair of markers had an effect, although dropping selected markers 5 and 7 was only significant at the nominal level $0{\cdot}10 > P > 0{\cdot}05$. This suggests that there must at least be two genetic effects on protein % linked to this chromosome, one prior to marker 5 (at 41 cM) in the selected marker set and one post marker 7 (at 54 cM).

Although not all traits showed evidence for QTLs in the exploratory analyses they were all analysed with multiple marker interval mapping methods. The analyses were carried out at 1 cM intervals. Three empirical thresholds were determined: the $P < 0{\cdot}05$ nominal level assuming one test and the thresholds for overall suggestive and significant linkage assuming 3 independent traits and 29 autosomal bovine chromosomes resulting in a total of 87 tests. For each trait the thresholds were obtained by 100 000 permutations. The estimated thresholds are given in Table 2.

There was no evidence of any effect in the analysis of fat yield. For milk yield, protein yield and fat % the maximum F value from the least squares interval analysis were 2·10, 1·99 and 2·34 and exceeded the nominal threshold value of $P < 0{\cdot}05$. For protein % the maximum F value was 2·89 and hence exceeded the threshold for suggestive linkage and came very close to the overall significance level derived assuming 29 independent chromosomes and 3 independent traits. The F-ratio curve for protein % is shown in Figure 1.

Table 2 *Empirical significance thresholds from permutation analysis for F ratios from least squares, single QTL analyses*

	Significance level		
Trait	Nominal, single trait	Suggestive, three traits	Suggestive, three traits
Milk yield	1·99	2·29	2·89
Fat yield	1·87	2·17	2·61
Protein yield	1·88	2·18	2·72
Fat %	2·07	2·39	3·01
Protein %	2·03	2·37	2·95

The results from analyses of protein % data within individual families are shown in Figure 2 for families for which the test statistic exceeded the nominal 0·05 level. The nominal significance of the maximum F value in these families varied from $P < 0{\cdot}025$ to $P < 0{\cdot}001$. The most likely position for a QTL was not very consistent across these families. In fact at the most likely position of the QTL from the across family analysis there was only a significant effect for grandsires 1 and 16 with an estimated substitution effect on protein % of 0·12 and 0·09, respectively. The

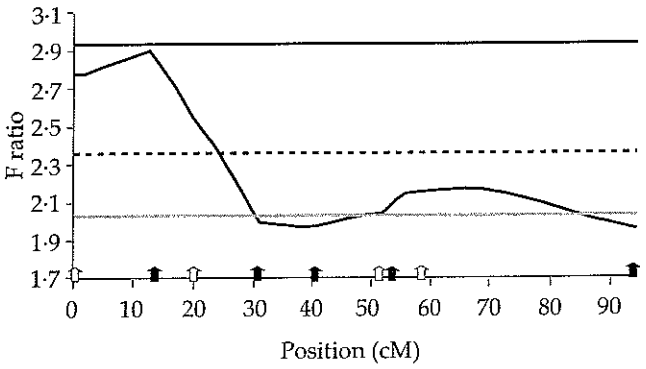

Figure 1 The F values through the linkage group from least-squares interval analysis fitting one QTL affecting protein %. The horizontal lines representing significance thresholds are $P_{nominal} < 0{\cdot}05$ (———), 'suggestive' linkage (- - - - - - -) and $P_{overall} < 0{\cdot}05$ (———). Positions of markers are shown by arrows, with filled arrows used for markers selected for the exploratory analyses.

other three families had their maximum test statistic towards the other end of the linkage group.

In order to test whether there was significant heterogeneity in the position estimates for protein % between families two log-likelihood ratios were calculated for each grandsire: one fitting a single QTL at the best location for that grandsire and one fixing the position of the QTL at the best overall position (obtained from a joint analysis of all 20 grandsires). The log likelihood for each grandsire is:

$$LR_i = n_i\,[\log_e \mathrm{RSS1}_i - \log_e \mathrm{RSS2}_i]$$

$\mathrm{RSS1}_i$ = residual sum of squares for grandsire i fitting only a mean; $\mathrm{RSS2}_i$ = residual sum of squares for grandsire i fitting a single QTL; n_i = number of sons for grandsire i.

The overall test was then calculated as the sum of the difference between the two tests for each grandsire, i.e.:

$$\mathrm{TEST} = \sum_{i=1}^{s} [LR_i \text{ (best location for grandsire } i) - LR_i \text{ (global location for grandsire } i)]$$

where s is the number of grandsires. This test statistic should approximately be distributed as χ^2 with 19 d.f. (for the 19 additional locations estimated). For protein %, the value of TEST = 32·2, giving a nominal $P < 0{\cdot}05$. This suggests that there may indeed be heterogeneity between families in the location of QTLs, possibly because different QTLs are segregating in different families.

The interval analysis fitting two QTLs at all combinations of positions was carried out for all

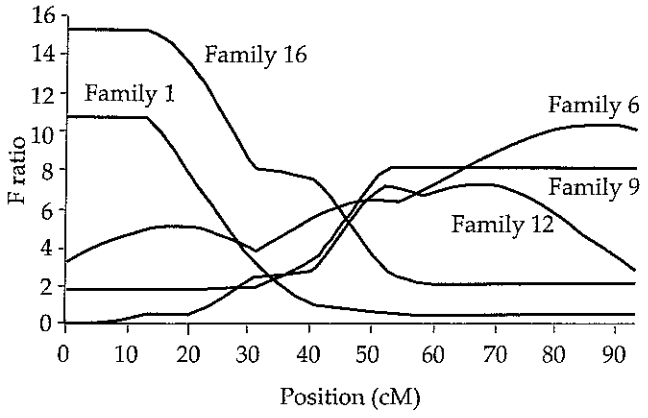

Figure 2 The F values through the linkage group from least-squares interval analysis fitting one QTL affecting protein % in individual families 1, 6, 9, 12 and 16.

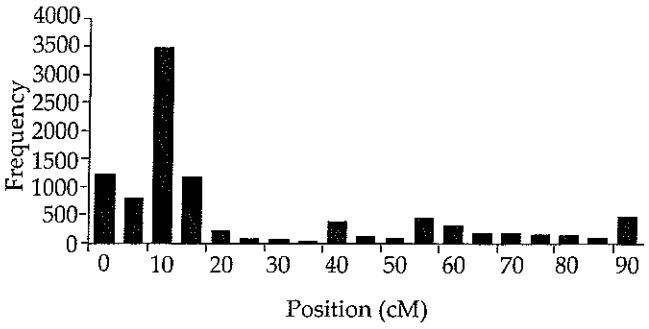

Figure 3 Distribution of best estimates of the position of a single QTL from least-squares interval analysis after 10 000 bootstrap resamples of protein % data.

traits and the results were compared to the one QTL analysis by an F test. For milk yield, fat % and protein % the two QTL model was better with nominal probabilities of $P \approx 0{\cdot}05$, $0{\cdot}05 > P > 0{\cdot}01$ and $0{\cdot}05 > P > 0{\cdot}01$, respectively and for the other two traits $P > 0{\cdot}05$. The estimated positions of the two QTLs were 5 and 55 cM for protein %. The best single QTL model for protein % accounts for 8·1% of the residual variance, inclusion of the second QTL increased the residual variance accounted for to 12·8%.

The 95% confidence intervals for location obtained by 10 000 bootstrap samples for each trait fitting one QTL covered in all cases almost the whole linkage group. The distribution of the maxima for protein % is presented in Figure 3. The major mode of the maxima is close to the best estimated position from the one QTL analysis (i.e. 13 cM) and about 65% of the estimates are within 20 cM of this position, but the remainder are spread across the whole chromosome.

From previous analyses there was an indication that there was more than one chromosomal region affecting protein %, so for this trait the bootstrap analysis was repeated fitting two QTLs. The 95% confidence interval for the first fitted QTL ranged from 0 cM to 55 cM and for the second QTL it ranged from 20 cM to 90 cM. The combinations of positions are presented as a two dimensional view in Figure 4. This figure is of course symmetric around the diagonal. Maxima appear at several combinations of positions and although one of these positions is often around 55 cM, this is not always the case.

Finally, we applied interval mapping via the approximate ML approach following Knott *et al.* (1996). In analyses fitting a single QTL, both fat % and protein % gave strong evidence for a significant effect at the proximal end of the linkage group, with

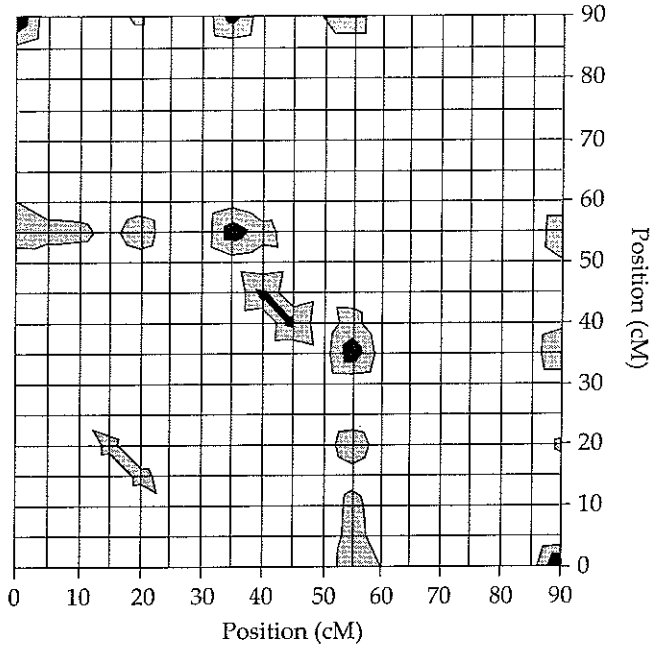

Figure 4 Distribution of best estimates of the joint position of two QTLs from least-squares interval analysis after 10000 bootstrap resamples of protein % data. Frequencies: black represents 400-600, grey represents 200-400, and white represents 0-200.

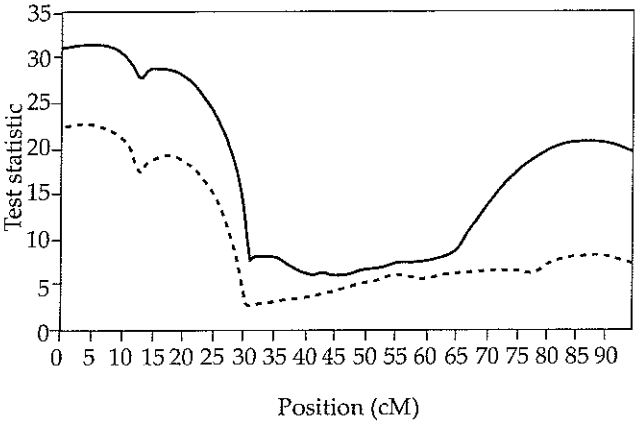

Figure 5 Log-likelihood test statistic through the linkage group from the approximate maximum likelihood analysis of a single QTL affecting protein %. The two curves show the log-likelihood curve for the test of a single QTL in the linkage group *versus* no QTL (————) and for the test of a single QTL in the linkage group plus an unlinked QTL *versus* only an unlinked QTL (--------).

maximum log-likelihood test statistics of 22·6 and 31·3 at 0 cM and 5 cM, respectively. These test statistics are for the alternative hypothesis of a linked QTL versus no QTL on the linkage group and give $P < 10^{-5}$ for both fat % and protein %, respectively when compared with a χ_2^2 distribution. The test statistic curve resembles the figures for the regression methods as can be seen in Figure 5 for protein %, but seems to be inflated at the ends of the linkage groups in relation to the least squares test statistic curves. Like the least squares interval mapping results, the analysis using ML gave the highest test statistic for protein %. For this trait the results of the ML analysis were compared with those from least squares. Using least-squares methods and fitting a single QTL, families 1 and 16 had a significant F ratio at the overall most likely position of a QTL (13 cM) when analysed separately (Figure 2). These families were therefore assumed to be heterozygous for a QTL at this position, the average substitution effect being 0·11%. In the ML analysis, the overall estimated frequency of heterozygous sires was 0·10 and sires of these families had an estimated probability of 0·9755 and 1·00, respectively, of being heterozygous for a QTL at 5 cM. The estimated substitution effect on protein % of the QTL was 0·13%. The sires of other families all had estimated probabilities of less than 0·01 of being heterozygous for a QTL at 5 cM.

The maximum likelihood analyses were then performed for two QTLs in the linkage group for the trait protein %. Computational limitations meant that the grid search could only be performed at 9 cM intervals for each pair of positions of the two QTLs and in addition the best locations for two QTLs from the regression analyses (5 and 55 cM) were examined. The likelihood was maximum at this latter pair of positions, where the log-likelihood ratio for this model compared with the best one QTL model was 12·1, a nominal probability of $0{\cdot}005 > P > 0{\cdot}001$. The overall probabilities of sires being heterozygous for these two QTLs were 0·10 and 0·82, respectively and their estimated substitution effects on protein composition were 0·14% and –0·02%, respectively.

Discussion

In this study we have explored the use of a range of analytical methods for dissecting quantitative trait locus effects in data from a dairy population structured as a number of half-sib families We have used these analyses to show the likely occurrence of at least one and probably more QTLs affecting milk protein % on bovine chromosome 6. The conclusion that genetic effects on this trait reside on chromosome 6 is not unique. Georges *et al.* (1995) reached a similar conclusion on a different sample from the Holstein population. Spelman *et al.* (1996) concluded that an effect was located on chromosome 6 in broadly the same set of data as we have analysed here. However, the main purpose of this study was to demonstrate the use of a structured analysis of

data of this type in order to develop as full a picture as possible of the QTL effects.

The basis of the methods we employ is the use of marker data to derive 'virtual markers' at chosen points through the genome. Analyses are then performed that are conditional on this virtual marker information. We have used relatively simple methods to reconstruct the likely sire gametes and derive virtual marker probabilities. Although more sophisticated methods may provide more information in some circumstances, we have previously shown that even complete knowledge of the true sire information may not greatly improve the power to locate QTLs (Knott *et al.*, 1996).

The procedure we employ is based firstly on use of exploratory analyses using regression on information at selected marker locations. This is followed by interval mapping approaches using regression and subsequently an approximate likelihood method. These methods can be complemented by use of permutation analysis and bootstrap analysis where appropriate and feasible to set significance thresholds and to derive confidence intervals.

The regression-based analyses are robust and fast to compute (facilitating the use of permutation and bootstrap analyses). The exploratory analyses provide a rapid method of giving an overview of effects associated with a particular linkage group and present no problem for setting significance thresholds (as the starting point is an independent test for each linkage group). Interval mapping by regression can provide more detailed information on the location of single QTL whilst retaining robustness. The ML approach allows more detailed inference on QTL effects and sire heterozygosity to be drawn with some penalty in the speed of computation, increased parameterization of the model and potential loss of robustness.

The results of the analyses are broadly consistent with one another. Only protein % gave much evidence of being influenced by loci on chromosome 6. For this trait the exploratory analyses suggested that at least two QTLs were segregating in the population and in fact hinted that there might be more than two. The least significant pair of the selected markers were 5 and 7, and dropping these from the analysis produced a near significant ($0{\cdot}10 > P > 0{\cdot}05$) reduction in fit. If this test had been significant, then the minimum number of separate QTLs in the linkage group needed to explain these results would have been three (one near or outside each of markers 2 and 9 and one near marker 5). If we accept that this test was not significant, then a minimum of two QTLs can explain the result (one near marker 4 and one near or outside marker 9).

The interval mapping analyses of protein % fitting two QTLs also suggested that at least two QTLs were segregating in the population. The regression and ML single and two QTL analyses also give similar test statistic surfaces, although as noted previously, the surface for the ML test tends to be relatively larger than the regression test statistic at the ends of the linkage group. This inflation of the test surface in the one QTL analysis occurs at the end of the linkage group where the marker information is rather weak. This raises the question of whether an effect unlinked to the chromosome under study could be inflating this test statistic in regions where marker information is scarce. We tested this possibility by fitting a model with one QTL linked to the chromosome and one unlinked QTL. This model proved to be a significant improvement over a model with just one linked QTL, the log-likelihood test statistic being 6·8 (2 d.f.), it was also a significant improvement over a model including only an unlinked QTL, the log-likelihood test statistic for the best position of the linked QTL being 22·6 (3 d.f.). Having included the additional QTL, the apparent magnification of the likelihood curve at the end of the linkage group was reduced (Figure 5). These results could be taken to demonstrate the presence of a major gene unlinked to the chromosome under study but we think it just as likely that it reflects failures of the underlying model, such as heteroskedastic within sire variances or non-normally distributed data, for which an unlinked QTL provides a partial explanation. Segregation analysis is very sensitive to failure of assumptions and so other potential causes of model failure should be examined very closely before accepting the explanation that major gene segregation is the cause. The least-squares based exploratory and interval analyses are much more robust to these sorts of data problems because they only use mean differences between marker genotypes in making inferences.

The results of the least-squares interval mapping are very similar to those of Spelman *et al.* (1996), who used only this analytical method and had slightly different data set with the same grandsire families and markers as used here, but incorporating more animals. Spelman *et al.* (1996) also had information about the genetic relationships between sire families and the actual length of the microsatellite alleles, which allowed them to make additional inferences about the inheritance of putative QTL alleles. Spelman *et al.* (1996) only found significant effects influencing protein % and were cautious in concluding the presence of a second QTL, but had exactly the same estimated position for the single

QTL analysis (13 cM) as well as similar estimated effects.

The several types of analysis used here point to there being possibly a second QTL at around 55 cM on the linkage group. This is very close to the position of the casein loci (marker 6 at 52 cM; Spelman *et al.*, 1996) which have previously been implicated in effects on fat percent (Bovenhuis and Weller, 1994). However, we should perhaps treat estimates of position and effect for this QTL with caution, particularly because of the hints of a third QTL from the exploratory analyses. A putative third QTL, or perhaps distributional problems with the data, or even the relatively small number of sires may explain the high estimated frequency of heterozygous sires (0·82) obtained from the maximum likelihood analysis which included two QTLs.

One difficult problem, which has yet to be fully resolved in QTL mapping, is that of the significance threshold. The exploratory analyses provide some means of addressing this as the initial test for an overall effect associated with a chromosome 6 can have a significance threshold determined solely by the number of chromosomes being examined. For a single QTL analysis by an interval mapping approach, significance thresholds can be set by simulation or by permutation (Churchill and Doerge, 1994). However, setting the test threshold for a second QTL once one has been detected remains problematical and requires further study.

A further problematical issue is that of multiple traits. Here we chose to set thresholds on the assumption that three independent traits were being analysed and we wanted to have an overall type I error of 5%. As the number of independent traits in a study increases, maintaining the study wide level of type I errors would lead to a gradually increasing rate of type II errors. Extending this argument *ad absurdam* might lead the very cautious to adopt a career wide significance threshold, resulting in a low type I error rate, but thresholds so stringent that no significant effects were ever located! In general then, a reasonable approach is to use a threshold appropriate for a genome scan of a single trait and accept an increased overall rate of type I errors over all traits studied.

It is clear that detection of a second QTL even in the relatively large sample analysed in this study, will always be challenging, and being able to conclude the three or more QTLs are segregating, will often be well nigh impossible. The problem is compounded to some extent by the fact that markers are not completely informative, so it becomes completely impossible to resolve closely linked QTL using least squares methods. In principle use of ML methods can allow such models to be analysed, but in practice the additional information that can be extracted by such methods is minimal. A better solution would be to add some more informative markers to regions of the map such as chromosome 6 where there is evidence for the presence of more than one QTL. Even with a dense map of highly informative markers resolution will be limited by the sample size and hence the availability of recombination events separating closely linked markers. Additional resolution may also be gained by a move towards analytical methods that take joint account of information from several correlated traits and such methods need to be developed for outbred population structures.

In conclusion, QTL mapping studies demand care in their analysis and interpretation of their results. This is underlined by the scale and cost of the studies and the potential cost of mistaken inferences, both in misguided breeding decisions and the cost of follow-up mapping studies. The strategy developed here is to use simple least-squares methods initially and follow these up by more focused and highly parameterized analyses as appropriate. The former analyses provide for robust exploratory analyses, whilst the latter analyses are capable of allowing more detailed interpretation of the data where their underlying assumptions are appropriate. This latter point is important as our results suggest that effects not associated with the linkage group can cause inflated test-statistics within the linkage group, especially where markers are low in information. We have used an approximate ML approach for the more focused analyses, and despite its relative simplicity this method has been shown to be effective for this data structure (Knott *et al.*, 1996; Elsen *et al.*, 1997). Full-blown maximum likelihood or Monte-Carlo sampling based approach, although more computationally demanding, could also be appropriate.

Acknowledgements

Wouter Coppieters, Michel Georges and their laboratory staff at the University of Liege are gratefully acknowledged for providing the genotypes for bovine chromosome 6. We also thank Johan van Arendonk for making the phenotypic data available. We acknowledge support for this work from the EU, in addition, CSH acknowledges support from BBSRC and MAFF and SAK acknowledges support from BBSRC and the Royal Society. We also thank an anonymous referee for suggesting simplifications and clarifications of the original text.

References

Andersson, L., Haley, C. S., Ellegren, H., Knott, S. A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfors-Lilja, I., Fredholm, M., Hansson, I., Hakansson, J. and Lundström, K. 1994. Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* **263:** 1771-1774.

Bovenhuis, H. and Weller, J. I. 1994. Mapping and analysis of dairy-cattle quantitative trait loci by maximum likelihood methodology using milk protein genes as genetic markers. *Genetics* **137:** 267-280.

Chatfield, C. and Collins, A. J. 1989. *Introduction to multivariate analysis.* Chapman and Hall, London.

Churchill, G. A. and Doerge, R. W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* **138:** 963-971.

Dekkers, J. M. C. and Dentine, M. R. 1991. Quantitative genetic variance associated with chromosomal markers in segregating populations. *Theoretical and Applied Genetics* **81:** 212-220.

Elsen, J. M., Knott, S. A., Le Roy, P. and Haley, C. S. 1997. Comparison between some approximate maximum likelihood methods for quantitative trait locus detection in progeny test designs. *Theoretical and Applied Genetics* **95:** 236-245.

Genstat 5 Committee. 1993. *GENSTAT 5 release 3 reference manual.* Clarendon Press, Oxford.

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., Sargeant, L. S., Sorensen, A., Steele, M. R., Zhao, X., Womack, J. E. and Hoeschele, I. 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139:** 907-920.

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

Haley, C. S., Knott, S. A. and Elsen, J. M. 1994. Multi marker approaches to quantitative trait loci in livestock. *Proceedings of the 45th meeting of the European Association of Animal Production, Edinburgh.*

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

Knott, S. A., Elsen, J. M. and Haley, C. S. 1994. Multiple marker mapping of quantitative trait loci in half sib populations. *Proceedings of the fifth world congress on genetics applied to livestock production, Guelph, vol. 21,* pp. 33-36.

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

Knott, S. A., Haley, C. S. and Thompson, R. 1991. Methods of segregation analysis for animal breeding data: parameter estimates. *Heredity* **68:** 313-320.

Kruglyak, L. and Lander, E. S. 1995. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57:** 439-454.

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

Lander, E. and Kruglyak, L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11:** 241-247.

Martinez, O. and Curnow, R. N. 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85:** 480-488.

Neimann-Sorensen, A. and Robertson, A. 1961. The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agriculturae Scandinavica* **11:** 163-196.

Ooijen, J. W. van. 1992. Accuracy of mapping quantitative trait loci in autogamous species. *Theoretical and Applied Genetics* **84:** 803-811.

Ron, M., Band, M., Yanai, A. and Weller, J. I. 1994. Mapping quantitative trait loci with DNA microsatellites in a commercial dairy cattle population. *Animal Genetics* **25:** 259-264.

Spelman, R. J., Coppieters, W., Karim, L., Arendonk, J. A. M. van and 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.

Van Raden, P. M. and Wiggans, G. R. 1991. Derivation, calculation and use of national animal-model information. *Journal of Dairy Science* **74:** 2737-2746.

Vilkki, H. J., Koning, D.-J. de, Elo, K., Velmala, R. and Mäki-Tanila, A. 1997. Multiple marker mapping of quantitative trait loci of Finnish dairy cattle by regression. *Journal of Dairy Science* **80:** 198-204.

Visscher, P. M. 1996. Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. *Journal of Heredity* **87:** 136-138.

Visscher, P. M. and Haley, C. S. 1996. Detection of putative quantitative trait loci in line crosses under infinitesimal genetic models. *Theoretical and Applied Genetics* **93:** 691-702.

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

Weller, J. I., Kashi, Y. and 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.

Whittaker, J. C., Thompson, R. and Visscher, P. M. 1996. On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77:** 23-32.

(*Received* 10 *July* 1997—*Accepted* 2 *April* 1998)