Theor Appl Genet (1996) 93:71–80

S. A. Knott · J. M. Elsen · C. S. Haley

Methods for multiple-marker mapping of quantitative trait loci in half-sib populations

Received: 5 June 1995 / Accepted: 19 January 1996

Abstract In this paper we consider the detection of individual loci controlling quantitative traits of interest (quantitative trait loci or QTLs) in the large half-sib family structure found in some species. Two simple approaches using multiple markers are proposed, one using least squares and the other maximum likelihood. These methods are intended to provide a relatively fast screening of the entire genome to pinpoint regions of interest for further investigation. They are compared with a more traditional single-marker least-squares approach. The use of multiple markers is shown to increase power and has the advantage of providing an estimate for the location of the QTL. The maximum-likelihood and the least-squares approaches using multiple markers give similar power and estimates for the QTL location, although the likelihood approach also provides estimates of the QTL effect and sire heterozygote frequency. A number of assumptions have been made in order to make the likelihood calculations feasible, however, and computationally it is still more demanding than the least-squares approach. The least-squares approach using multiple markers provides a fast method that can easily be extended to include additional effects.

Key words QTL mapping · Genetic mapping · Animal breeding · Half-sibs

Communicated by C. Smith

S. A. Knott (✉)[1] · C. S. Haley[2]
INRA Station de Génétique Quantitative et Appliquée Jouy-en-Josas, France

J. M. Elsen
INRA Station d'Amélioration Génétique des Animaux, Castanet-Tolosan, France

[1] *Present address*: Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK
[2] *Present address*: Roslin Institute (Edinburgh) UK

Introduction

Genetic maps of the major livestock species based on molecular genetic markers provide the tools to begin the mapping of some of the loci underlying quantitative traits – so called quantitative trait loci or QTLs (Geldermann 1975). These studies will not only provide insight into the control of these economically important traits, but will also allow the enhancement of breeding programmes through the application of marker-assisted selection. The mapping of a QTL may also ultimately allow it to be cloned for more detailed study in the laboratory.

The cost of genotyping animals for a large number of markers means that large samples will be expensive to achieve and this puts great emphasis on the efficient use of the data collected. A number of authors have considered designs for the analysis of data from half-sib populations (e.g. Neimann-Sorensen and Robertson 1961; Soller and Genizi 1978; Geldermann et al. 1985; Dentine and Cowan 1990; Weller et al. 1990; Le Roy and Elsen 1995). The drawback of these methods is that they use information from a single marker at a time. No marker will have a heterozygosity of unity and recent studies (Georges et al. 1995; Barendse et al. 1994) using highly informative markers, such as multi-allelic microsatellites, have reported average heterozygosities in the range 0.5 to 0.7, so for any given marker some sires will be homozygous and thus uninformative. Single-locus analyses waste information and results from other situations suggest that there will be a potentially greater problem of introducing bias into the estimated location of QTLs (Haley et al. 1994). If several markers of different information content are linked to the QTL the marker which produces the strongest evidence for a QTL may be that which is most informative rather than that which is closest. An additional problem of using only single markers is that of separately estimating the position and effect of any detected QTL. The least-squares methods that have been proposed (e.g.

Neimann-Sorensen and Robertson 1961; Soller and Genizi 1978; Geldermann et al. 1985) can not distinguish between an effect tightly linked to the marker and a larger effect more loosely linked. Maximum-likelihood methods (Weller 1986, 1990) can potentially estimate both effects, but estimates are generally poor using only a single marker (Weller 1986; Knott and Haley 1992 a,b) and location is relative to the marker (i.e. the QTL could be either side of the marker).

Interval mapping (Lander and Botstein 1989) has been found to be more powerful than the use of single markers for the analysis of populations derived from a cross between inbred lines and to provide more accurate estimates of the position and effect of a QTL (Knott and Haley 1992 b; Darvasi et al. 1993). The application of interval-mapping approaches to data from outbred populations is not straightforward and can be computationally demanding. Furthermore, because markers are not completely heterozygous, the information content varies from interval to interval depending upon the markers flanking that interval. This presents the same potential bias as with single-marker analyses, in that there may be a bias towards locating a QTL in the most informative interval rather than the correct one (Knott and Haley 1992 a; Haley et al. 1994). The problem of bias in the location of a QTL due to variation in the information content of markers can potentially be overcome by simultaneous use of all of the markers in a linkage group (Haley et al. 1994).

Georges et al. (1995) present a maximum-likelihood approach to QTL detection for use in half-sib populations. Their approach makes use of information from all markers in a linkage group simultaneously, but analyses families separately. They consider all possible reconstructions of each sire's gametes and, hence, the analyses are computationally demanding.

In the present paper we demonstrate the use of all markers in a linkage group for the analysis of data from half-sib populations. This requires the extension and modification of previously developed methods, firstly to calculate transmission probabilities in two-generation half-sib families and, secondly, to allow the linkage phase to differ from family to family . Methods using information from all markers have been developed using both a least-squares and an approximate maxmum-likelihood approach. The aim is to provide relatively fast methods for preliminary analysis of the entire genome which would be used in conjunction with alternative approaches which may give a more detailed description of the situation. These methods are compared with the traditional single-marker ANOVA approach by the analysis of simulated data.

Methods

The methods used assume that trait data have been collected from the half-sib progeny of a number of unrelated sires. These data could be milk-records on females or they could be weighted breeding values on a number of half-sib sons estimated from data collected on their daughters (these situations correspond to the 'Daughter' and 'Granddaughter' designs discussed by Weller et al. 1990). Sires are assumed to be randomly mated to unrelated dams and each dam to have only a single progeny. Marker data are available on the sires and their offspring and may or may not be available for the dams. We assume that the order of markers in a linkage group and the distance between them is known.

The basic philosophy is similar to that in interval mapping (Lander and Botstein 1989), which has previously been applied to the analysis of data from crosses between inbred and outbred lines (Haley and Knott 1992; Martinez and Curnow 1992; Haley et al. 1994). In this approach, for given positions (e.g. 1 cM intervals) through a linkage group, the probability of an offspring inheriting one or other of its parents' gametes at that position is calculated conditional on its marker genotype. As applied to a half-sib design, where little or no information on QTL/marker linkages can come from the dam, it is of interest only to calculate these probabilities for the sire gamete. Once these probabilities have been calculated, they can be incorporated into either a least-squares or a maximum-likelihood analysis. The analysis proceeds sequentially and will be considered in this order: firstly, inferring the marker alleles inherited from the sire by each progeny and reconstruction of the sire gametes for the markers; secondly, calculating the probabilities of inheriting each sire gamete in each position for each offspring; and, thirdly, using this information in either a least-squares or an approximate maximum-likelihood analysis.

Marker inheritance and sire gamete reconstruction

Each marker in each sire-family is considered in turn. Markers for which a sire is homozygous are uninformative and are omitted from consideration. For markers which are heterozygous in the sire it may be possible to determine which allele a progeny has inherited (if the progeny possesses only one of the two sire alleles). If dam genotype information is available, this will increase the frequency of the sire allele inherited by a progeny being determined unequivocally.

Once informative markers have been identified and their inheritance determined, the gametes for each sire can be reconstructed for the linkage group under consideration. This is done simply by considering, in turn, each pair of adjacent markers for which the sire is heterozygous. Progeny in which the allele inherited from the sire can be determined at both loci are ascertained and the linkage phase is taken as that which minimises the number of recombination events in the sire. If both phases are equally likely, one is selected at random. This is repeated for each pair of adjacent heterozygous markers to reconstruct the two gametes for each sire. An example is given in Fig. 1.

This approach does not use all the information available from the half-sib progeny, but is expected to perform well for large half-sib families and is much faster than a complete analysis using all the half-sibs and linked markers simultaneously. In practice, when only a single data set is being analysed, it may be preferable to use a method of reconstruction that uses more information from the data.

Conditional probabilities of sire gamete inheritance

Throughout this paper it is assumed that there is no interference in recombination events and so Haldane's mapping function applies. The probabilities for each progeny inheriting the two sire gametes are calculated for fixed positions through the linkage group conditional upon their marker genotypes. On the assumption that the sire reconstruction is correct, and for each offspring using only markers for which the allele inherited from the sire is known unequivocally, these probabilities are the same as for a backcross situation which have been presented by Martinez and Curnow (1992). The probabilities depend only upon the alleles inherited at the two nearest informative markers flanking the position under consideration and the recombination between the markers and this position. In fact, as the conditional probabilities sum to unity, only that for the first sire gamete need be calculated. For any position, the markers used to

A sire has the following genotype:
Aa BB Cc dd EE Ff
The sire has 100 half-sib progeny. The following can be determined (dams have not been genotyped):
No. offspring known to inherit A and C or a and c = 6
No. offspring known to inherit A and c or a and C = 17
No. offspring known to inherit C and F or c and f = 15
No. offspring known to inherit C and f or c and F = 10.
Therefore the reconstructed sire gametes are:
Gamete 1: A c f
Gamete 2: a C F.
Two of the sire's offspring have the following genotypes at the loci of interest:
HS 1: AA cc ff
HS 2: aa Cc ff.
We wish to calculate the probability of inheriting the allele from sire gamete 1 for a QTL (Q) at 30 cM from marker A:

Offspring	Formula for conditional probability	Probability
HS 1	$(1-r_{AQ})(1-r_{QC})/(1-r_{AC})$	0.97
HS 2	$r_{AQ}(1-r_{QF})/r_{AF}$	0.50

r_{ij} is the recombination frequency between loci i and j

Fig. 1 Example calculation of conditional probabilities. Consider the situation with six markers at 20-cM intervals. Each marker has two alleles which are segregating at equal frequency in the population.

calculate these conditional probabilities will vary from sire to sire and from progeny to progeny within a sire. For some individuals a chosen position may be outside the last informative marker in the linkage group, in which case the conditional probabilities depend only on the single nearest informative marker. If all markers in a linkage group are uninformative in an individual the conditional probabilities would be 0.5 for both gametes at all positions in the linkage group. An example is given in Fig. 1.

Least squares analysis

For a given position the conditional probabilities of the offspring inheriting the first gamete of the sire provide an independent variable on which the trait score can be regressed. For a single sire this would provide an estimate of the substitution effect (Falconer 1989) for a heterozygous QTL at that position. For the simultaneous analysis of several sires the regression must be nested within sires. This is because not all sires will be heterozygous for any QTL and, for those that are, the linkage phase between the QTL and the sire gamete which has been designated as first will vary from sire to sire. The between-gamete within-sire regression term, with degrees of freedom equal to the number of sires, is compared to the residual mean square to provide an F ratio test for the presence of a QTL. The analysis is repeated at fixed locations (e.g. every 1 cM) throughout the genome and the position maximising the F ratio is considered to be the most likely location for any QTL. These F-ratio statistics can also be plotted against the position for which it was calculated to provide a curve displaying evidence for the presence of a QTL through the linkage group.

Approximate maximum-likelihood analysis

Full maximum likelihood for this type of problem is computationally demanding. Thus we make a number of approximations which make the problem much more tractable. Firstly, we assume that the effect of the QTL is relatively small, so the effect of its segregation on the distribution within groups of animals inheriting the same sire gamete is unimportant. Secondly, we assume that there are only two QTL alleles segregating at this locus in the population and their frequency is the same in the different groups of dams mated to each sire. Thirdly, we have previously shown that between-family genetic variation is a potential source of bias if not accounted for (Knott and Haley 1992 a). Here we assume that the progeny group sizes are sufficiently large such that the effect can be removed by adjusting the data from each half-sib into a deviation from the mean of that group.

With these assumptions the likelihood for a QTL in a given position requires only three parameters: the frequency of sires homozygous at the QTL (p), the substitution effect (Falconer 1989) of the QTL (α), and the residual variance within groups of progeny inheriting one or other of the sire gametes (σ_w^2). The likelihood is:

$$
\begin{aligned}
L = \prod_{i=1}^{s} \Bigg\{ & p \prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_w^2)^{1/2}} \exp\left(\frac{-z_{ij}^2}{2\sigma_w^2}\right) \\
& + \frac{(1-p)}{2} \prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_w^2)^{1/2}} \left[m_{ij} \exp\left(\frac{-(z_{ij}-\alpha/2)^2}{2\sigma_w^2}\right) \right. \\
& \left. + (1-m_{ij}) \exp\left(\frac{-(z_{ij}+\alpha/2)^2}{2\sigma_w^2}\right) \right] \\
& + \frac{(1-p)}{2} \prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_w^2)^{1/2}} \left[m_{ij} \exp\left(\frac{-(z_{ij}+\alpha/2)^2}{2\sigma_w^2}\right) \right. \\
& \left. + (1-m_{ij}) \exp\left(\frac{-(z_{ij}-\alpha/2)^2}{2\sigma_w^2}\right) \right] \Bigg\}
\end{aligned}
$$

where z_{ij} is the adjusted record for the jth half-sib offspring of the ith sire and m_{ij} is the conditional probability that offspring j inherits gamete 1 from sire i at the position being considered. Sire i has n_i offspring.

The accuracy of this approximation has been compared to that of more complicated likelihoods by Elsen et al., in preparation. Using a single fully informative marker for the same HS design considered here they compared the above approximation with three others (Demenais et al. 1990; Boichard et al. 1990; Knott et al. 1992) which involved more parameters, describing both the between-family variance and the within-sire distribution. They found that, in terms of power, the approximation used here was a good as the best of the alternative approximations. The estimates of the substitution effect and variance were close to the simulated values. For a QTL of smaller effect, however, the frequency of heterozygous sires at the QTL was overestimated.

As for the least-squares approach described above, this likelihood is optimised at fixed locations along the chromosome, and compared with the null-hypothesis likelihood that there is no QTL. The position maximising the difference in likelihoods gives the most likely location of the QTL. For likelihood optimisation a quasi-Newton routine (E04JAF) from the NAG library was used (Numerical Algorithms Group 1990).

Simulation study

In order to evaluate the multiple-marker methods and to compare their performance with traditional single-marker methods, an analysis of simulated data is used.

Single-marker analysis

The principles of using analysis of variance to detect linked QTLs have been previously described (Neimann-Sorenson and Robertson 1961; Weller et al. 1990). For each marker in turn, informative (i.e. heterozygous) sires and offspring in which the allele inherited from the sire can be identified are selected and a test is provided by an

F test of the ratio of the between-marker allele within-sire mean square to the residual mean square. Even for markers that are expected to have the same information content, chance variation in the markers that are heterozygous in the sire and those informative in the offspring will cause the degrees of freedom of both the numerator and denominator of this test to vary from marker to marker. To allow for this, the probability of the F ratio for each marker was determined and the most significant selected as providing the best estimate of the closest marker to any QTL.

Test statistic under the null hypothesis

When considering a single location for the QTL, the distribution of the test statistic when no QTL is segregating for the least-squares (LS) methods is asymptotically F, with the degrees of freedom being the number of sires included for the numerator and the total number of offspring minus twice the number of sires for the denominator. For full maximum likelihood (ML) twice the difference in the likelihood with and without a QTL should follow a χ^2 distribution with two degrees of freedom (for the two additional parameters, the frequency of heterozygous sires and for the substitution effect at the QTL). For these analyses, however, we are using an approximate maximum-likelihood method. Furthermore, for all methods a number of tests are being carried out (at 1-cM points through the linkage group for the multiple-marker methods or at each marker for the single-marker method) which are not independent and hence the distribution of the test statistic under the null hypothesis is difficult to determine theoretically. Therefore we will use simulation to arrive at an empirical distribution for the tests and data structure we are using.

Data were simulated for 20 sires each with 100 half-sib progeny. Each individual was composed of a 100-cM chromosome with markers at either 10 cM, 20 cM or 50 cM intervals. Markers had either two or four alleles segregating at equal frequency in the population. A sire has a 0.5 probability of being heterozygous for a marker with two alleles at equal frequency and, if the sire is heterozygous, there is a 0.75 probability of determining which sire allele an offspring has inherited if dam genotype information is available and a 0.5 probability if dam-genotype data is not. For a marker with four alleles at equal frequency these probabilities are 0.75, 0.9375 and 0.75, respectively. A phenotype was simulated for each individual for a polygenic trait with a heritability of 0.24. For each situation 10 000 simulations and analyses were carried out in order to obtain suitable significance thresholds for the LS methods and 1000 were performed for the ML approach. In one set of replicates dam-marker genotype information was utilised, in the other it was ignored. For the multiple-marker LS analyses, degrees of freedom may differ between replicates because some sires may be completely uninformative and these are dropped from the analysis. Hence, as for the single markers, direct comparison of F ratios may not be possible and instead the probability of each F ratio was used. In this way all analyses can contribute to a single distribution of the test statistic. For ML all sires are retained in the analysis. To aid comparison, however, rather than using the likelihood-ratio test statistic the probability of this statistic coming from a χ^2 distribution with two degrees of freedom was used.

Over the 10 000, or 1000, replicates the significance level which would give a whole-chromosome type-I error of 5% and 1% was determined (i.e. the level which 5% or 1% of replicates, respectively, would be expected to exceed by chance somewhere on the chromosome if no QTL were segregating).

Simulations with a QTL

When a QTL is segregating in the population, we are interested in both the power of the proposed methods to detect the QTL and in the estimates of its location and effect. To investigate these properties 100 replicates were simulated with the genomes and population described above with the addition of a QTL. Various alternative situations were considered for the QTL. A QTL with an additive effect of 1.09 within-QTL genotype standard deviations between homozygotes was simulated at position 25 cM (for the 10 and 50-cM spaced markers) or 30 cM (for the 20-cM spaced markers), which places the QTL half-way between the markers in each case. Additionally, in the genome with markers spaced at 20-cM intervals, the same effect QTL was placed 40 cM from the end of the chromosome which places the QTL at a marker.

In addition to the situations described above, data were simulated where markers on one chromosome varied in their expected information content. Six markers at 20-cM intervals were simulated with the first three having two alleles segregating at equal frequency and the last three having four alleles. Two situations were considered, one with an additive QTL of 1.09 within-QTL genotype standard deviations between homozygotes simulated to be 30 cM from the end of the chromosome (i.e. flanked by low-information markers) and another with the same effect QTL simulated at 50 cM (i.e. flanked by one low- and one high-information marker).

Results

Null hypothesis

Under the null hypothesis using single markers the probability of the most significant F ratio selected from the markers in a linkage group for each analysis was not related to the number of sires used (i.e. those heterozygous for a given marker). Additionally, on average, the degrees of freedom were not inflated above those expected (e.g. for a marker with two alleles, we would expect the average numerator and denominator degrees of freedom to be 10 and 730, respectively, when dam information is used, as 50% of sires and 75% of their offspring are expected to be informative). This suggests that there is no tendency for the markers with a higher number of informative sires to have the most significant test statistic when no QTL is segregating. For multiple-marker LS there was much less variation in the number of sires used over analyses (as few sires were uninformative for all markers in a linkage group) and no evidence that the significance of the F ratio was related to the number of sires in the analysis. Thus for both methods the use of the probability of a single test to derive the multiple-test significance threshold seems reasonable.

The simulated empirical thresholds are given in Table 1. For the single-marker analyses these are close to, although slightly higher than, those that would be obtained on the assumption that the tests at individual loci were independent using the Bonferroni adjustment to the thresholds for multiple tests. Despite a larger number of tests being performed in the multiple-marker approaches (every cM rather than at each marker), the significance thresholds tended to be higher (i.e. to be significant the probability of the F ratio of χ^2 under the null hypothesis has to be less extreme) particularly where markers were close together. The thresholds for ML were much higher (i.e. less extreme) than for LS with multiple markers and with widely spaced markers were even higher than expected for a single test.

Sire gamete reconstruction

With the dense map and informative markers, errors in reconstruction were rare. The worst situation was with the 50-cM spaced markers with two alleles and without dam information, when 6% of informative sires over all replicates were incorrectly reconstruted. Including dam information halved this number.

Power

An example of the results produced by plotting the test statistic against the chromosomal position are shown in Fig. 2. The same data were analysed with all three methods. The test statistic is the probability of the likelihood ratio or the F ratio under the null hypothesis and, hence, the most likely location for the QTL is the position giving the lowest probability. The markers or

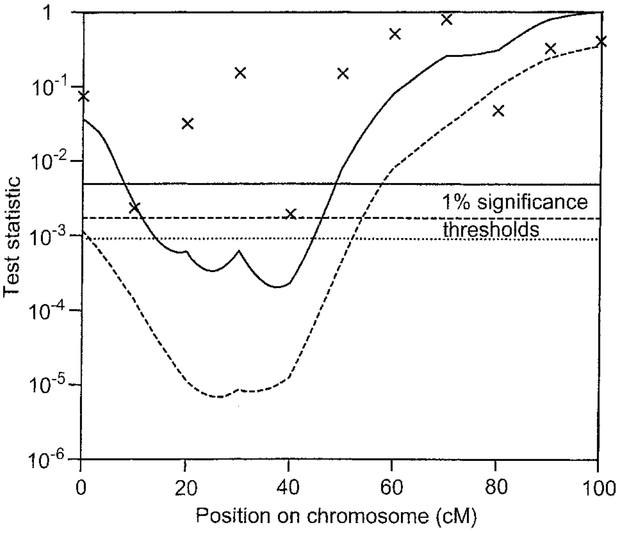

Fig. 2 An example of the test statistics obtained from the three methods for a single set of data. Markers were located at 10-cM intervals and each one had two alleles segregating at equal frequency. Dam-marker genotypes were not included in the analyses. The QTL was additive in effect with 1.09 within-QTL genotype standard deviations between homozygotes and was located at 25 cM. The results and empirical 1% threshold for multiple markers with ML are shown by *solid lines* and for LS by *dashed lines*. The results from the single-marker analyses are shown by a cross and the threshold by a *dotted line*

Table 1 Empirical 0.01 and 0.05 significance thresholds

Method	Marker alleles	Significance threshold	10-cM interval		20-cM interval		50-cM interval	
			Dams	No dams	Dams	No dams	Dams	No dams
Single marker	2	0.01	0.00081	0.00090	0.00182	0.00163	0.00378	0.00378
		0.05	0.00482	0.00447	0.00893	0.00898	0.01757	0.01703
	4	0.01	0.00111	0.00081	0.00156	0.00173	0.00396	0.00362
		0.05	0.00539	0.00483	0.00910	0.00894	0.01806	0.01714
Multiple-marker LS	2	0.01	0.00164	0.00172	0.00302	0.00274	0.00410	0.00386
		0.05	0.01019	0.01123	0.01422	0.01529	0.02183	0.02068
	4	0.01	0.00120	0.00111	0.00145	0.00191	0.00277	0.00278
		0.05	0.00685	0.00724	0.00948	0.01109	0.01585	0.01576
Multiple-marker ML	2	0.01	0.00347	0.00490	0.00683	0.01015	0.01199	0.01589
		0.05	0.04416	0.04047	0.04543	0.04877	0.07249	0.07050
	4	0.01	0.00525	0.00731	0.00688	0.00537	0.01039	0.01303
		0.05	0.03363	0.03154	0.04729	0.04759	0.06237	0.06094

Based on 10000 replicate simulations of each situation for LS analyses and 1000 for ML. Presented as the probability of the relevant F ratio or χ^2

locations below the relevant 1% significance threshold are significant at the 1% level. The multiple-marker approaches give similar curves which dip around the simulated position of the QTL (25 cM). The curve obtained using ML is higher than that obtained using LS, but the significance threshold is also higher, so a shorter region of the chromosome gave a significant test using ML rather than LS. For the single-marker method tests are carried out for each marker. The markers flanking the simulated QTL location do not always have the lowest probability because, to some extent, the probability depends on the number of heterozygous sires at the marker.

For all methods and structures the power was high when compared against the 0.05 threshold. Differences between the methods are more marked at intermediate powers and, hence, we focus on the 0.01 threshold. The percentages of analyses significant at the empirical 0.01 threshold are shown in Table 2. The use of multiple markers increases the power in all situations. The increase in power from the use of multiple markers was greatest when the markers were close together and the power was intermediate. As expected, the use of dam-genotype information increases the power. Using multiple markers, ML and LS gave similar powers.

Table 2 also indicates that if the power is not already high, there is a useful increase in power on going from 20-cM to 10-cM marker spacing, which is not the case for data from an inbred-line cross (Darvasi et al. 1993).

The situation with the QTL placed half-way between markers is least favourable for the single-marker analyses. Table 2 also gives the power when the QTL is located at a marker. This causes an increase in power when the power is not already very high. The advantage in power of using multiple markers, however, is maintained in this situation.

The power obtained from the replicates that had at least one sire incorrectly reconstructed was not significantly different from the power in the remaining replicates. When the data was analysed using LS with the correct reconstruction of the sire gametes the power did not alter substantially. For the 20-cM spaced markers a change in power was observed only for the situation with two alleles when dam information was not used, and then gave an increase of only 2% when considering the 1% significance level. For the 50-cM spaced markers, small changes in power were observed (up to 4%). There was no trend in the change in power, and for two of the situations the power decreased with the correct reconstructions.

Parameter estimates

Location

Mean estimates of position and of the empirical standard deviation of the position estimate are shown in Table 3. Estimates from the single-marker analyses have been converted to a cM-position for comparative purposes. For the more informative situations the mean estimates of position were reasonable. For the wider spaced markers with only two alleles, and when dam information was omitted, mean estimates are biased. When the power of detection is low the mean estimate for the QTL location is expected to be biased towards the centre of the chromosome, as the estimates for location when a QTL is not detected should be distributed throughout the genome.

The standard deviation of the position estimate was generally less from the multiple-marker analyses, only half the value of that from the equivalent single-marker analysis in extreme cases. In general the standard deviation was decreased by the use of dam-genotype information and by decreasing the marker interval for all methods. Using multiple markers, in most of the situations considered, the variances of the location estimates were lower with ML than LS.

Table 4 shows that when the QTL was located at a marker all methods provided unbiased estimates of the location. The variances of the location estimates for single-marker LS are closer to those from the multiple-marker approaches and for all methods the variances are lower than when the QTL was located between markers.

Table 2 Percentage of replicates significant at the empirical 0.01 threshold

Method	Marker alleles	10-cM interval		20-cM interval				50-cM interval	
		QTL at 25 cM		QTL at 30 cM		QTL at 40 cM		QTL at 25 cM	
		Dams	No dams	Dams	No dams	Dams	No dams	Dams	No dams
Single marker	2	80	57	67	41	81	46	34	21
	4	92	91	95	86	96	89	76	56
Multiple-marker LS	2	95	89	93	74	92	80	37	25
	4	99	98	97	96	97	98	85	74
Multiple-marker ML	2	92	84	87	75	91	78	42	30
	4	94	95	97	93	98	96	81	72

Based on 100 replicate simulations of each situation. The QTL was additive with two alleles at equal frequency and 1.09 within-QTL genotype standard deviations between the mean effect of the homozygous genotypes.

Table 3 Mean estimates of the position of the QTL (and their empirical standard deviation) for QTLs positioned between markers

Method	Marker alleles	10-cM interval		20-cM interval		50-cM interval	
		Dams	No dams	Dams	No dams	Dams	No dams
Single marker	2	25.5 (15.7)	30.2 (20.3)	28.8 (15.4)	34.4 (20.3)	30.5 (32.5)	36.5 (35.4)
	4	26.0 (9.6)	26.3 (11.9)	31.0 (11.8)	30.6 (15.2)	24.0 (26.1)	27.5 (27.9)
Multiple-marker LS	2	24.6 (7.3)	25.7 (13.4)	30.8 (13.8)	34.8 (18.1)	31.9 (25.1)	39.0 (32.6)
	4	25.4 (4.8)	26.2 (7.9)	29.8 (8.1)	29.7 (9.2)	23.2 (14.3)	25.1 (18.9)
Multiple-marker ML	2	25.1 (7.8)	25.2 (9.5)	31.4 (12.4)	33.7 (16.2)	28.7 (22.4)	37.0 (30.5)
	4	26.1 (9.4)	25.6 (7.0)	29.5 (7.2)	29.2 (8.9)	24.6 (13.0)	25.9 (17.6)

Based on 100 replicate simulations of each situation. The QTL was additive with two alleles at equal frequency and 1.09 within-QTL genotype standard deviations between the mean effect of the homozygous genotypes. The simulated positions were 25 cM, 30 cM and 25 cM for the 10-, 20- and 50-cM intervals respectively

Table 4 Mean estimates of the position of the QTL (and their empirical standard deviation) for QTLs at a marker

Method	Marker alleles	Dams	No dams
Single marker	2	39.2 (13.3)	42.6 (18.3)
	4	40.4 (5.7)	40.8 (11.3)
Multiple-marker LS	2	38.5 (9.7)	38.7 (15.3)
	4	40.7 (6.2)	40.4 (6.0)
Multiple-marker ML	2	39.2 (8.2)	40.8 (12.5)
	4	40.2 (4.2)	40.7 (4.8)

Based on 100 replicate simulations of each situation. A 20-cM marker map was used with the QTL at 40 cM (at a marker). The QTL was additive with two alleles at equal frequency and 1.09 within-QTL genotype standard deviations between the mean effect of the homozygous genotypes

Table 5 gives the situations with markers of varying information content. The estimate of location is biased when using only single markers, especially when the true QTL location is flanked by markers of differing information content, whereas the use of multiple markers provides an unbiased estimate.

Analysing the data with correct sire gamete reconstruction made almost no difference to the mean estimate and standard deviation for QTL location. Considering just those replicates where at least one sire was incorrectly reconstructed, for most situations the mean alteration in location was less than 1 cM, the only exception being the situation with 50-cM spaced markers with two alleles when dam information was included and where the mean location in the 37 replicates that had at least one sire incorrect changed from 31.1 to 32.7 cM.

QTL variance

The variance explained by the QTL under the ML approach can be estimated by $(1-p)\alpha^2/4$ where $(1-p)$ is the frequency of heterozygous sires and α is the substitution effect at the QTL. This is the variance explained by the within-family segregation of the sire's QTL alleles and, assuming the QTL is at the estimated location, should equal $\sigma^2_{QA}/4$ where σ^2_{QA} is the additive variance at the QTL. The estimates obtained are given in Table 6. On average, the frequency of heterozygous sires was overestimated by the ML method, the overestimation being greater when using less informative markers. The substitution effect, on the other hand, was generally slightly underestimated for these QTLs. For example, using makers with two alleles at 50-cM intervals gave a mean estimate for the heterozygote frequency of 0.66 with a standard deviation of 0.35, and for the substitution effect of 0.54 with a standard deviation of 0.24. Thus, the two biases approximately cancel giving a reasonable estimate for the QTL variance explained by the markers (see Table 6). The inclusion of dam-marker genotypes and an increase in the information content of the markers did not have a consistent effect on the variance estimates when analysed with ML. There is some indication, however, that with wider-

Table 5 Mean estimates of the position of the QTL (and their empirical standard deviation) with mixed-information markers

Method	QTL at 30 cM		QTL at 50 cM	
	Dams	No dams	Dams	No dams
Single marker	32.8 (18.8)	36.4 (25.0)	55.2 (12.4)	56.8 (16.0)
Multiple-marker LS	28.3 (14.4)	30.0 (18.5)	49.3 (12.7)	50.6 (17.1)
Multiple-marker ML	27.5 (12.0)	30.4 (17.7)	49.7 (10.8)	51.2 (14.3)

Based on 100 replicate simulations of each situation. Markers were simulated at 20-cM intervals with the first three markers having two alleles segregating at equal frequency and the last three having four. The QTL was additive with two alleles at equal frequency and 1.09 within-QTL genotype standard deviations between the mean effect of the homozygous genotypes

Table 6 Mean estimates of the sire marker-associated QTL variance (and empirical standard deviation)

Method	Marker alleles	10-cM interval		20-cM interval		50-cM interval	
		Dams	No dams	Dams	No dams	Dams	No dams
Single marker	2	0.040 (0.014)	0.049 (0.019)	0.033 (0.014)	0.037 (0.015)	0.020 (0.012)	0.022 (0.016)
	4	0.035 (0.013)	0.037 (0.014)	0.031 (0.010)	0.032 (0.012)	0.019 (0.007)	0.020 (0.009)
Expected values		0.030		0.025		0.014	
Multiple-marker LS	2	0.029 (0.012)	0.025 (0.011)	0.021 (0.008)	0.016 (0.007)	0.011 (0.007)	0.008 (0.005)
	4	0.032 (0.011)	0.031 (0.011)	0.029 (0.010)	0.027 (0.010)	0.018 (0.007)	0.016 (0.006)
Multiple-marker ML	2	0.038 (0.012)	0.038 (0.014)	0.038 (0.014)	0.039 (0.016)	0.036 (0.021)	0.038 (0.025)
	4	0.035 (0.013)	0.035 (0.013)	0.039 (0.013)	0.040 (0.014)	0.039 (0.015)	0.039 (0.017)
Expected values		0.037		0.037		0.037	

Based on 100 replicate simulations of each situation. The QTL was additive with two alleles at equal frequency and 1.09 within-QTL genotype standard deviations between the mean effect of the homozygous genotypes

spaced markers both of these factors cause a decrease in the empirical standard deviation of the estimates.

For the LS analyses one measure of QTL variance can be obtained by considering the difference in residual MS between a model fitting the QTL and one in which it is omitted. If the QTL allele inherited from the sire was known this difference in MS would also equal $\sigma^2_{QA}/4$. For single-marker analyses, the allele inherited at the markers is known for all individuals included in the analysis, but the QTL variance explained will be decreased because of recombination between the marker and the QTL (denoted r). The expected variances are decreased by a factor of $(1-2r)^2$. Estimates are shown in Table 6. After accounting for the effects of recombination, the QTL variance explained by the selected marker was, on average, greater than that expected. This is because the selected marker will be the one that explains the most variance (i.e. there is a bias due to selection of the most significant marker). Both increasing the number of alleles at the markers and including dam information cause a decrease in the empirical standard deviation of the estimate because of an increase in the number of individuals included in the analyses.

Table 6 also gives this variance estimate from the LS analyses using multiple markers. This is expected to be lower than $\sigma^2_{QA}/4$ because for a given location of the QTL we have only the probability of an offspring inheriting each of the sire's alleles, not the actual allele inherited. The variance estimate is decreased, therefore, by a function of the recombination rates between the estimated location of the QTL and the flanking informative markers for each individual, giving expected values between those for ML and single markers. The mean values obtained are less than when using single markers because the effect of selecting the best location is much less than selecting the best marker, as the estimates at neighbouring locations are more highly correlated using multiple markers simultaneously. The effect of using the incorrect sire gamete reconstruction on the estimates was negligible.

A different estimate of the QTL variance could be obtained from the mean within-sire variance obtained from the substitution effect estimated for each sire. Alternatively the difference in mean squares could be adjusted using the probability of inheriting the sire alleles and the substitution effect.

Discussion

The methods presented here illustrate the use of half-sib data for the detection of QTLs. Approaches using multiple markers are advantageous giving both greater power and an improved estimate of the QTL location, particularly when markers vary in information content. A simplified likelihood has been used which enables the whole genome to be scanned rapidly. This likelihood approach did not perform better than the multiple-marker LS approach in terms of power and has the disadvantages of being much slower to compute, in particular if it is extended to include additional effects (such as additional QTLs or environmental factors). The LS approach used here may also be less affected by departures from normality than ML and does not require the assumption that only two alleles are segregating at the QTL, which may be required to make ML computationally tractable. Thus it may be more robust, and hence preferable to ML, for livestock populations under selection. One drawback of the LS approach is that estimates of the effect of the QTL are not available directly (although an indication is possible from the analysis of the sire substitution effects), but its speed and simplicity allow rapid scanning of the genome. With an accurate estimate of the location of the QTL and the potential of high power when informative markers are used, this method could be followed by the use of a computationally more demanding one, such as ML, on a restricted region of the genome enabling additional parameters (including the effect and frequency of the QTL) to be estimated.

Dam-genotype information

In practice we are unlikely to have marker genotypes for most dams. However, the analyses illustrate the maximum improvement that could be obtained using marker information from the dams. When dam genotypes are not known we could use information from the offspring about marker-allele frequency in the dams and include this when obtaining probabilities of inheritance of the sire alleles. This would complicate the analysis and the improvement would not be great. Knowledge of the dams' alleles would be most useful when the highest proportion of offspring have the same genotype as their heterozygous sire. This occurs when two of the marker alleles are at intermediate frequencies. In this situation inclusion of the dam-allele frequencies has no effect because the two allele frequencies are approximately equal in the dams and, hence, inclusion of this information does not change the probability of inheritance of the sire alleles.

Sire-gamete reconstruction

For each sire only a single reconstruction of the sire gamete has been considered. Using the simple method suggested, with a high number of half-sib progeny per sire, the correct reconstruction was obtained frequently. Additionally, incorrect reconstructions had little effect on the power and parameter estimates, presumably because incorrect reconstruction occurred in areas of low information and hence influence the results very little. With a low number of half-sibs the reconstruction will not be as good, which leads to less-accurate parameter estimates. Methods considering all possible reconstructions of the sire gametes (for example, that proposed by Georges et al. 1995) may be preferable in this situation, but they would also perform less well with few half-sibs.

Single-marker analyses

For single-marker analyses, when a QTL was simulated, the use of the most significant *F* ratio as a criterion to select the 'best' marker tends to pick markers with, on average, a higher than expected number of informative sires (with six markers, each with two alleles, the average number of sires for the marker with the most significant *F* value was 10.7 when dam information was used, significantly different from the expected value of 10). For example, two markers at equal distance from the QTL would be expected to have the same *F* ratio (with the same number of informative offspring per sire), but the one with more informative sires, and hence degrees of freedom, would be more significant. This bias can lead to markers not flanking the QTL to be chosen if they involve more sires than those closer. It is not clear, however, what criteria if any, might be more appropriate for selection of the best single marker. Alternative criteria have been investigated – e.g. using the reduction in the residual mean square caused by fitting the marker or the *F* ratio itself. These criteria varied in the power they gave and in their ability to estimate the location of the QTL but there was no clearcut 'best' method over all simulations studied. For the analyses presented, the criterion used here gave high power compared with the others.

Null hypothesis

For all methods, multiple simulations are required in order to obtain the empirical significance thresholds. In practice a permutation test (Churchill and Doerge 1994) with the real data would be used. The probabilities of inheriting either sire allele would have to be obtained only once for the whole genome and then these can be permuted against the phenotypes. This would take account of the correct marker structure and avoids the problem encountered here for multiple-marker LS where the degrees of freedom differed between replicates. The LS method, being rapid, makes such simulations practical.

Multiple QTLs

We have ignored the problem of multiple QTLs. For the multiple-marker methods, as with similar approaches, it would be possible to carry out multi-dimensional searches fitting two or more QTLs. Alternatively, the approaches proposed by Jansen (1993) and Zeng (1993), fitting marker co-factors to account for additional QTLs, could be implemented. In the analysis of populations derived from inbred lines the use of marker co-factors has been shown to be advantageous, reducing the residual variance and, hence, increasing power and avoiding biases due to linked QTLs. In the half-sib population the inclusion of marker co-factors will be less beneficial as, although the residual variance may be decreased, linked QTLs are less of a problem because the population is not in complete linkage disequilibrium.

In conclusion, the use of multiple markers is advantageous for half-sib populations as it is for outbred line crosses. The LS method provides a fast and relatively powerful method of harnessing the information in multiple markers.

Acknowledgements This work was initiated whilst S.A.K. and C.S.H. were the recipients of Institut National de la Recherche Agronomique (France)/Agricultural and Food Research Council (UK) fellowship awards. S.A.K. and C.S.H. are very grateful for the hospitality they received during their stays at INRA Station d'Amélioration Génétique des Animaux and Station de Génétique Quantitative et Appliquée. Additional financial support came from the Biotechnology and Biological Sciences Research Council (S.A.K. and C.S.H.), the Ministry of Agriculture, Fisheries and Food (C.S.H.) and a Human Capital and Mobility Project supported by the European Commission (S.A.K, J.M.E. and C.S.H.). We thank the referees and editor for helpful suggestions to improve the manuscript.

References

Barendse W, Armitage SM, Kossarek LM, Shalom A, Kirkpatrick BW, Ryan AM, Clayton D, Li L, Neibergs HL, Zhang N, Grosse WM, Weiss J, Creighton P, McCarthy F, Ron M, Teale AJ, Fries R, McGraw RA, Moore SS, Georges M, Soller M, Womack JE, Hetzel DJS (1994) A genetic linkage map of the bovine genome. Nature Genet 6:227–235.

Boichard D, Elsen JM, Le Roy P, Bonaiti, B (1990) Segregation analysis of fat content in Holstein × European Friesian cattle. In: Hill WG, Thompson R, Woolliams J (eds) Proc 4th World Congr Genet Appl Livest Prod, Edinburgh, XIV:167–170

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

Darvasi A, Vinreb A, Minke V, Weller JI, Soller M (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134:943–951

Demenais FM, Murigane C, Bonney GE (1990) Search for faster methods of fitting the regressive models to quantitative traits. Genet Epidemiol 7:319–334

Dentine MR, Cowan CM (1990) An analytical model for the estimation of chromosome substitution effects in the offspring of individuals heterozygous at a segregating marker locus. Theor Appl Genet 79:775–780

Falconer DS (1989) Introduction to quantitative genetics, 3rd edn. Longman, UK

Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers. I. Methods. Theor Appl Genet 46:319–330

Geldermann H, Pieper U, Roth B (1995) Effect of marked chromosome sections on milk performance in cattle. Theor Appl Genet 70:138–146

Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR, Zhao X, Womack JE, Hoeschele I (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139:907–920

Haley CS, Knott SA (1992) A simple method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Haley CS, Knott SA, Elsen JM (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136:1195–1207

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135:205–211

Knott SA, Haley CS (1992a) Maximum-likelihood mapping of quantitative trait loci using full-sib families. Genetics 132:1211–1222

Knott SA, Haley CS (1992b) Aspects of maximum-likelihood methods for the mapping of quantitative trait loci in line crosses. Genet Res 60:139–151

Knott SA, Haley CS, Thompson R (1992) Methods of segregation analysis for animal breeding data: a comparison of power. Heredity 68:299–311

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

Le Roy P, Elsen JM (1995) Numerical comparison between powers of maximum-likelihood and analysis of variance methods for QTL detection in progeny test designs: the case of monogenic inheritance. Theor Appl Genet 90:65–72

Martinez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480–488

Neimann-Sorensen A, Robertson A (1961) The association between blood groups and several production characteristics in three Danish cattle breeds. Acta Agric Scand 11:163–196

Numerical Algorithms Group (1990) The NAG Fortran Library Manual – Mark 14. NAG Ltd, Oxford, UK

Soller M, Genizi A (1978) The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. Biometrics 34:47–55

Weller JI (1986) Maximum-likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627–640

Weller JI (1990) Experimental designs for mapping quantitative trait loci in segregating populations. In: Hill WG, Thompson R, Woolliams J (eds) Proc 4th World Congr Genet Appl Livest Prod, Edinburgh, XII:113–116

Weller JI, Kashi Y, Soller M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative loci in dairy cattle. J Dairy Sci 73:2525–2537

Zeng ZB (1993) Theoretical basis of precision mapping of quantitative trait loci. Proc Natal Acad Sci USA 90:10972–10976