Heredity **81** (1998) 48–54 *Received 13 August 1997, accepted 23 January 1998*

Simple multiple-marker sib-pair analysis for mapping quantitative trait loci

SARA A. KNOTT*† & CHRIS S. HALEY‡

†Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT and ‡Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, U.K.

Haseman & Elston (1972) developed a simple and robust method for detecting quantitative trait loci (QTL) by their linkage to a single marker using information from sib-pairs. The method involves the regression of the squared difference between the phenotypic scores onto the proportion of alleles at the marker which are identical by descent. The availability of genetic maps of marker loci makes it possible to extend this method to incorporate information from several marker loci. Here we show that by considering identity by descent from the two parents separately, a simple method can be obtained to use information from multiple markers to estimate the proportion of alleles identical by descent for a QTL in any given location. Considered this way, the method can also be very simply extended to allow for differences in recombination frequencies between the sexes and more complicated relationships between individuals. We show by simulation that this method is as powerful as alternative least squares approaches using multiple markers (Fulker *et al*., 1995) and provides more accurate estimates. It is also much easier to implement.

Keywords: least squares, QTL mapping, sib-pairs.

Introduction

Haseman & Elston (1972) developed a simple sib-pair method for detecting QTLs in outbred populations where families of full-sibs were available. In this approach the squared difference between the phenotypic scores of two full-sibs (Y) is regressed onto the proportion of alleles for which the two sibs are identical by descent (ibd) at a particular marker (π). Haseman & Elston (1972) showed that with a QTL which was a recombination fraction θ from the marker, the expectation of the regression coefficient, β, was:

$$\beta = -2(1-2\theta)^2\,\sigma_g^2,$$

where σ_g^2 is the additive genetic variance contributed by the QTL. Fulker & Cardon (1994) extended the sib-pair method to use information from pairs of flanking markers and, subsequently, Fulker *et al*. (1995) further extended it to use multiple markers. Their approach is to estimate the proportion of alleles ibd in two sibs at each marker and to use this information to obtain an estimate of the proportion of alleles ibd at any given location (π_q):

$$\widehat{\pi}_q = \hat{\alpha} + \beta_1\widehat{\pi}_1 + \beta_2\widehat{\pi}_2 + \ldots + \beta_n\widehat{\pi}_n,$$

where n is the number of markers in the linkage group; the βs are obtained from a series of simultaneous equations: $\mathbf{C} = \mathbf{V}\boldsymbol{\beta}$, where $\mathbf{V}$ is the expected variance matrix of the proportion ibd at the markers and $\mathbf{C}$ contains the expected covariances between the proportion ibd at the markers and the location being considered; the $\hat{\pi}$s ($\widehat{\pi}_1, \ldots, \widehat{\pi}_n$) are the estimates of the proportion ibd at the markers; and $\hat{\alpha}$ is estimated from the $\boldsymbol{\beta}$ coefficients and the mean ibd states at the markers.

Fulker *et al*. (1995) illustrated that the multiple marker approach removed biasses in location observed with the flanking markers, as expected, and in some situations increased power. The main drawback to this approach, however, is the difficulty of extension to include differences in male and female recombination fractions and more complicated pedigree structures.

The approach also has a potential bias in that markers at which the ibd state cannot be determined are assumed to have the expected value based on a single marker, i.e. if the ibd state of alleles from both parents cannot be determined a value of 0.5 is used, and if the ibd state for alleles from only one parent cannot be ascertained this parent contributes

*Correspondence. E-mail: s.knott@ed.ac.uk

a value of 0.25 to the ibd state for this marker. If flanking markers are available, however, the probability of being ibd may differ from the single marker value. One example, where this use of expected values would lead to the largest deviation from the correct probability, would be in a situation where one marker was heterozygous for both parents and offspring and its flanking markers were either both completely ibd or completely not ibd. With a marker heterozygous for the same alleles in parents and both sibs, either the alleles inherited by both sibs are ibd from both parents or from neither parent. If no additional information were available then this marker would have an estimated ibd state of 0.5. If both flanking markers are ibd from both parents, however, it is much more likely that the marker in between is also completely ibd rather than completely not ibd. In which case the estimated ibd state for this marker is greater than 0.5.

Method

We follow the basic approach adopted by Fulker *et al*. (1995). For a sib-pair we estimate the proportion of alleles ibd for a QTL at a known position conditional on the ibd state observed at the markers ($\widehat{\pi_q}$). This is carried out for fixed positions through the linkage group. The squared difference between the phenotypic scores of two full-sibs (Y) is then regressed onto $\widehat{\pi_q}$ at each position. As we approach a QTL, the recombination fraction, θ, between the location of the QTL and the position being considered, is reducing and the estimate of β increasing, and at a QTL $\beta = -2\sigma_g^2$. Therefore we are interested in large negative values of β. At each location a t-statistic ($\hat{\beta}/\mathrm{SE}(\hat{\beta})$) is calculated and the position at which this is minimized is selected as the best estimate for the location of any QTL. We implement the calculation of π_q differently from Fulker *et al*. (1995), however, in a way that is more readily extendible to a mix of full- and half-sibs, which allows differences in recombination between sexes to be accommodated, and is faster to run.

We start by noting that the number of alleles at a locus ibd in a pair of sibs (i.e. 0, 1 or 2) can also be thought of as whether the alleles inherited from the sire are ibd and whether the alleles from the dam are ibd. Then the pair of sibs will have two alleles ibd if both the sire and the dam alleles are ibd, one allele if either the sire or the dam allele is ibd and no alleles ibd if neither the sire nor the dam alleles are ibd. Considered in this way, with data on codominant markers from both parents and offspring, for one parent (either sire or dam) markers are either completely informative or they are completely uninformative. There are no partially informative markers when considering one parent at a time. However, because markers may be informative from one parent but uninformative from the other, if the parents are considered together a marker may be considered partially informative. In this method we use this observation and, for a QTL at a known position, estimate the probability that the two sibs are ibd from each parent separately ($\widehat{\pi_{q\mathrm{s}}}$ and $\widehat{\pi_{q\mathrm{d}}}$ for the sire and dam, respectively). These estimates are then combined to estimate the overall probability that the sibs are ibd for the QTL ($\widehat{\pi_q} = (\widehat{\pi_{q\mathrm{s}}} + \widehat{\pi_{q\mathrm{d}}})/2$).

Considering the gametes from the two parents separately, and taking the dam as an example, for each gamete $\widehat{\pi_{q\mathrm{d}}}$ is calculated for a known position between two markers. $\widehat{\pi_{q\mathrm{d}}}$ is calculated conditional on $\pi_{1\mathrm{d}}$ and $\pi_{2\mathrm{d}}$, these being the proportions of alleles which are ibd from the dam at the nearest informative markers flanking the QTL position (i.e. $\pi_{1\mathrm{d}}$ and $\pi_{2\mathrm{d}}$ will either be 0 or 1), and the recombination fractions between the position of the QTL and the markers, as shown in Table 1. The calculations shown in Table 1 assume no interference and under this assumption it is only the two flanking informative markers that provide information on $\pi_{q\mathrm{d}}$. Thus for a QTL in a fixed position for any one gamete (from the sire or dam), in any one pair of sibs, only two markers are required to calculate $\widehat{\pi_{q\mathrm{d}}}$ or $\widehat{\pi_{q\mathrm{s}}}$. However, the markers used may differ between the dam and the sire gametes (i.e. information from up to four markers may be used for a pair of sibs) and also between different pairs of sibs. Some positions towards the end of a linkage group may not be flanked by two informative markers, the outer marker being uninformative for one or both parents in some sib-pairs. In this case the probability that the sib-pair are ibd from the particular parent is calculated conditional on the nearest informative marker internal to the position, using formulae given in Table 1.

Note that for the situation with two fully informative (from both parents) markers and no sex difference in recombination frequencies this approach is algebraically equivalent to the method given by Fulker & Cardon (1994). Compared with Fulker *et al*. (1995) this method will be the same within regions containing only fully informative markers, with the variance of the ibd state within these markers being 0.125. The complete analysis proceeds as follows.

Table 1 Probabilities of identity by descent from the sire (or dam) at a QTL conditional on flanking markers or adjacent single marker and recombination frequencies

ibd state at flanking markers		Probability of ibd at QTL
π_1	π_2	$(\widehat{\pi_{qs}})$
1	1	$((1-\theta_1)^2+\theta_1^2)((1-\theta_2)^2+\theta_2^2)/((1-\theta)^2+\theta^2)$
1	0	$((1-\theta_1)^2+\theta_1^2)(1-\theta_2)\theta_2/((1-\theta)\theta)$
0	1	$(1-\theta_1)\theta_1((1-\theta_2)^2+\theta_2^2)/(1-\theta)\theta)$
0	0	$4(1-\theta_1)\theta_1(1-\theta_2)\theta_2/((1-\theta)^2+\theta^2)$
1	—[a]	$((1-\theta_1)^2+\theta_1^2)$
0	—	$2(1-\theta_1)\theta_1$

π_1, π_2, numbers of alleles ibd from the sire (or dam) at the first and second marker, respectively.
$\widehat{\pi_{qs}}$, estimated probability of the QTL being ibd from the sire (or dam), calculated assuming no interference in recombination (Haldane mapping function).
θ_1, θ_2, θ, recombination fractions between the first marker and the QTL, the second marker and the QTL and between the two markers, respectively, in the sire (or dam).
[a] Uninformative (i.e. for a position towards the end of a linkage group which is not flanked by two informative markers for that parent in that sib-pair).

1 For each pair of sibs identify markers in the linkage group that are informative for identity by descent for the two parents considered separately.
2 For a QTL at a fixed position in the linkage group calculate $\widehat{\pi_{qd}}$ and $\widehat{\pi_{qs}}$ conditional on the nearest two flanking markers that are informative for the appropriate parent (or the nearest single marker if the position is not flanked by two informative markers) and the appropriate recombination frequencies using the formulae in Table 1. Calculate $\widehat{\pi}_q = (\widehat{\pi_{qs}} + \widehat{\pi_{qd}})/2$.
3 Regress the squared phenotypic difference between the sib-pairs (Y) onto $\widehat{\pi}_q$ for the chosen position and calculate the t-statistic for the regression coefficient $\hat{\beta}$.
4 Repeat from 2 for chosen positions through the linkage group (e.g. 1 cM intervals). Select the position for which t is minimized and compare with the appropriate significance threshold. Estimate σ_g^2 as $-\hat{\beta}/2$.

Using this approach, splitting information from the sire and from the dam, there is one combination of alleles at a marker for which the ibd state is known but the origin from the sire or from the dam is not known. That is when both parents are heterozygous with the same genotype and one of the offspring is homozygous and the other heterozygous. The alleles inherited from one parent are ibd and those from the other parent are not. We cannot tell, however, from which parent the ibd alleles have been inherited. The approach outlined above assumes this to be an uninformative situation and omits this marker for this pair of full-sibs. An alternative approach, however, would be to consider two possible scenarios: (i) that the alleles inherited from the sire were ibd and those from the dam were not; and (ii) vice versa. This will obviously complicate the calculation of the probabilities of ibd as the sire and dam can no longer be treated independently. One way to calculate the required probabilities would be to calculate the probabilities of the different marker situations (i.e. considering the two different scenarios given above for any such locus) for both parents together and all markers flanked by fully informative markers (informative in the sire and dam) or ends of chromosomes. Then the probabilities of the sibs being ibd at a given location for each marker situation can be calculated and the two probabilities multiplied together and summed over the different situations. With this method it is still possible to take account of male and female recom-

bination differences. This complete method will bring the average of the sire and dam ibd state down to a value of 0.5 at this type of marker where both parents and one offspring have the same heterozygous genotype and the other sib is homozygous. In the simpler method described above, however, the probability of ibd will depend on the flanking markers. The worst situation will be when this type of marker falls between two at which the ibd state is completely known and the alleles inherited by the two full-sibs are either ibd from both parents or neither parent. This situation will not be common as it requires a recombination event in one parent on both sides of the marker.

Simulation

We use simulation to explore the properties of this simple multiple-marker method in comparison with the method proposed by Fulker *et al*. (1995). We also compare the complete form of the method proposed above. The comparisons between the methods are expected to be affected by marker density and heterozygosity, with the largest differences being when few alleles are segregating at the markers.

All simulations considered a linkage group 120 cM in length. Seven markers were equally spaced (20 cM apart), with two alleles at equal frequency. In all cases recombination with no interference was simulated. Where a QTL was simulated it had two alleles at equal frequency with a heritability of 0.5 (i.e. its segregation explained 50% of the total variance). To investigate the effect of QTL location, sets of replicate simulations were run. A single QTL was simulated in each set of replicates, with the position of the QTL varying between sets. In different sets, the position of the QTL was varied from 0 to 20 cM and 40 to 60 cM in 5 cM steps (i.e. 0, 5, 10, etc.) and 20 to 40 cM in 2 cM steps (i.e. 20, 22, 24, etc.). Thus, 19 sets of replicates covering the region 0 to 60 cM were used. QTLs with smaller effect (explaining 5% and 10% of the total variance) were also investigated. An additional situation was considered with marker spacing as before but with eight alleles with equal frequency segregating at all markers. The QTL was simulated at location 50 cM.

All populations considered 1000 pairs of sibs and their parents (except for the situations with smaller QTL effects where 10 000 sib-pairs were considered). For each situation 500 simulations were performed.

The analyses proceeded as described above for the multiple-marker method. For simplicity and comparison across methods all sib-pairs were included even when all markers in the linkage group were uninformative. The estimated position of any QTL was taken as that at which the t-statistic was minimized.

Significance thresholds

For a test at a single position or marker, the distribution of $\hat{\beta}/\mathrm{SE}(\hat{\beta})$ is expected to conform to a t-distribution. Many correlated tests at positions through the linkage group will be performed, however, and so thresholds need to be obtained empirically. With real data we would propose the use of a permutation test (Churchill & Doerge, 1994). In this test the ibd states are permuted with the phenotypic differences and the resulting data set reanalysed. Over multiple permutations this gives an empirical distribution of the test statistic under the null hypothesis of no QTL being present. For this study simulations were performed with the marker set-up as described above but with no QTL. For each combination of parameters 1000 simulations were performed and the 0.05 whole linkage group threshold for $\hat{\beta}/\mathrm{SE}(\hat{\beta})$ determined for each of the three methods.

Results

Significance thresholds

We report $-t$ values throughout to make the results and curves shown more directly comparable with those usually associated with interval mapping.

The significance thresholds for the entire linkage group are shown in Table 2 for the three methods. These were derived by selecting the highest $-t$ value from each analysis. We also looked at the distribution of test statistics for single locations and the thresholds obtained for the location 0 cM are given in Table 2. The results from other locations were similar. The equivalent value obtained from a standard t-distribution would be 1.65. We also looked at the mean and variance of the $-t$ values at positions along the chromosomes (data not shown). There was no evidence of any trends moving from the end to the centre of the linkage group or moving from a marker to positions between markers.

Comparative power

Figure 1 gives the power observed for the three methods for the situations considered with markers with only two alleles. All methods gave similar power, although the complete version using all information was always at least as good as the others

Table 2 Significance thresholds obtained empirically for the three methods. 5% values for a single location and for the whole chromosome are given

	Two alleles at the marker		Eight alleles at the marker	
Method[a]	Chromosomal 5%	Single location 5%[b]	Chromosomal 5%	Single location 5%[b]
Simple	2.37	1.62	2.52	1.63
Complete	2.39	1.60	2.55	1.63
FCC	2.37	1.57	2.56	1.64

[a] Simple and complete are the methods proposed in the text, FCC is the approach of Fulker *et al.* (1995).
[b] Values obtained for location 0 cM are given.

(except at 24 cM, where the simpler method proposed here was 1% higher). Although, as expected, power was higher when the QTL was simulated at a marker and outside the terminal intervals (i.e. in the more central intervals), the difference between the methods was not affected by QTL location. The set of simulations where eight alleles were simulated at the markers gave 96% power for all methods.

Estimates of position and σ_g^2

Results for the QTL location are given in Fig. 2. The mean of the best location for the QTL is expected to be biassed towards the centre of the chromosome unless there is 100% power. This is because of the boundary effect at the end of the linkage group and because type I errors produce an average position estimate at the centre of the linkage group. Within an interval there seems to be some bias towards the markers and hence the terminal part of an interval shows less bias than the central part. The methods are similar, with a tendency for that proposed by Fulker *et al.* (1995) to be most biassed (and with the largest variance of the best location estimate) and the approach described here using all information to be least biassed (with the smallest variance of estimates over the simulations). In the situation with the more informative markers (each with eight alleles) for all three

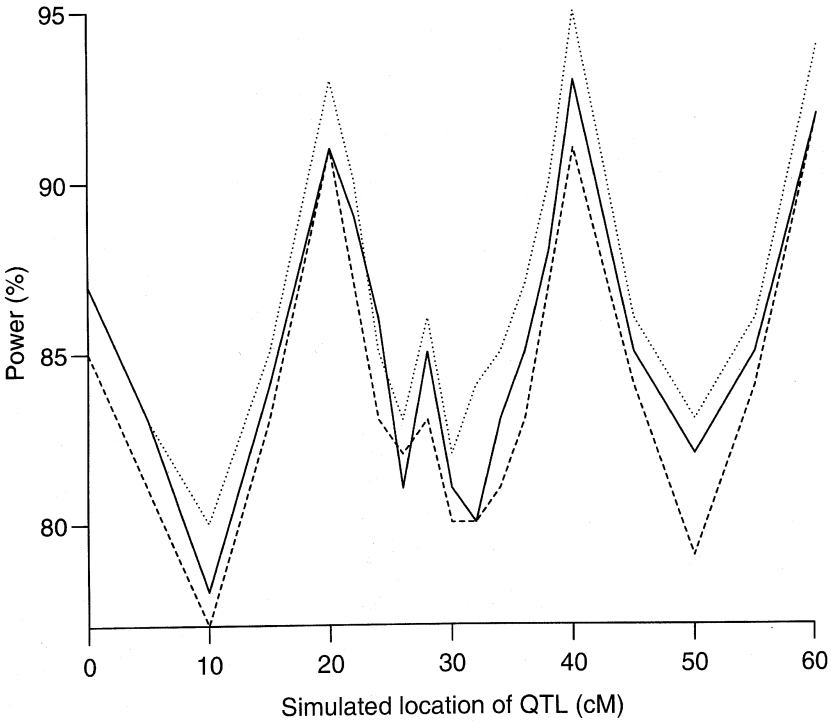

Fig. 1 Power observed for the three methods for QTLs simulated at different locations in the linkage group. The results from the simple approach are shown by a solid line, the complete approach by a dotted line and the approach of Fulker *et al.* (1995) by a dashed line.

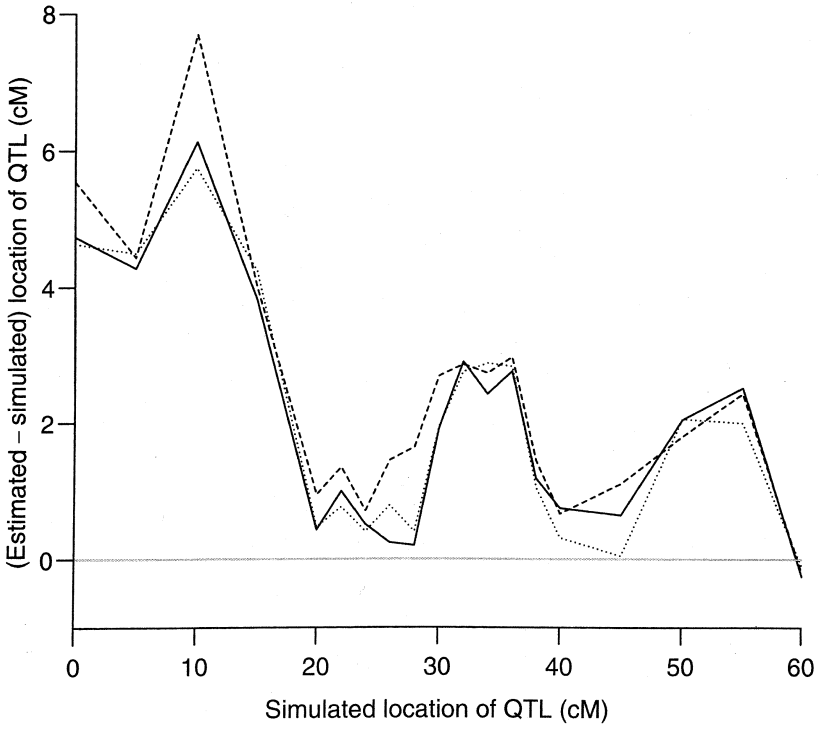

Fig. 2 The bias in estimated location of the QTL for the three methods for QTLs simulated at different locations in the linkage group. The results from the simple approach are shown by a solid line, the complete approach by a dotted line and the approach of Fulker *et al.* (1995) by a dashed line. The standard deviations of the location estimates vary between 9 cM at central markers and 20 cM at location 10 cM.

methods the estimate of the QTL location was, on average, at the simulated position of the QTL (with standard deviation of 10 cM). With the smaller QTL effects, the estimates of location were more biassed because of the reduced power. The differences between methods, however, were similar to that observed for the larger QTL effect. An alternative approach is to take the mean test statistic over all runs at each location. When considered in this way, the highest test statistic for all methods was the location of the simulated QTL, or at most 1 cM away.

Figure 3 gives the mean estimates for the QTL variance. The expected value is 0.5. The QTL variance estimates are consistently highest for the complete approach described here and intermediate for that of Fulker *et al.* (1995). In the situation where markers had eight alleles all methods gave an mean estimate over the replicate simulations of 0.51 (with standard deviation 0.13). With QTLs of smaller effect the overestimate of the QTL variance was greater, with the methods ranking as for the larger QTL effect.

Discussion

The three approaches give very similar results in the simulation study and, as expected, are most similar with the more informative markers. The location of the QTL affects the power and parameter estimates for the three methods in a similar way. The estimate for the variance of the QTL is generally overestimated for all methods. The selection of the highest variance estimate obtained in the linkage group (which by definition must be positive) explains part of this overestimate and disappears in the situation with more informative markers. It is not clear that one method is preferable on these grounds alone.

In all situations all families were included so comparison was made on exactly the same data. In practice it might be predicted to be better to omit families with no informative markers in a linkage group, as these are not informative about the presence of a QTL and if the QTL is segregating within them, they will increase the residual variance. In a simulation study, however, using the simple approach, we found that any selection of families, for example requiring at least one informative marker, or at least one informative marker in each parent, did not improve power or the estimates (results not shown). In fact a loss of power was observed and an associated overestimate of the QTL variance, especially when the selection meant a large decrease in the number of families included in the analysis.

The methods presented here all use information from multiple markers and will have the same

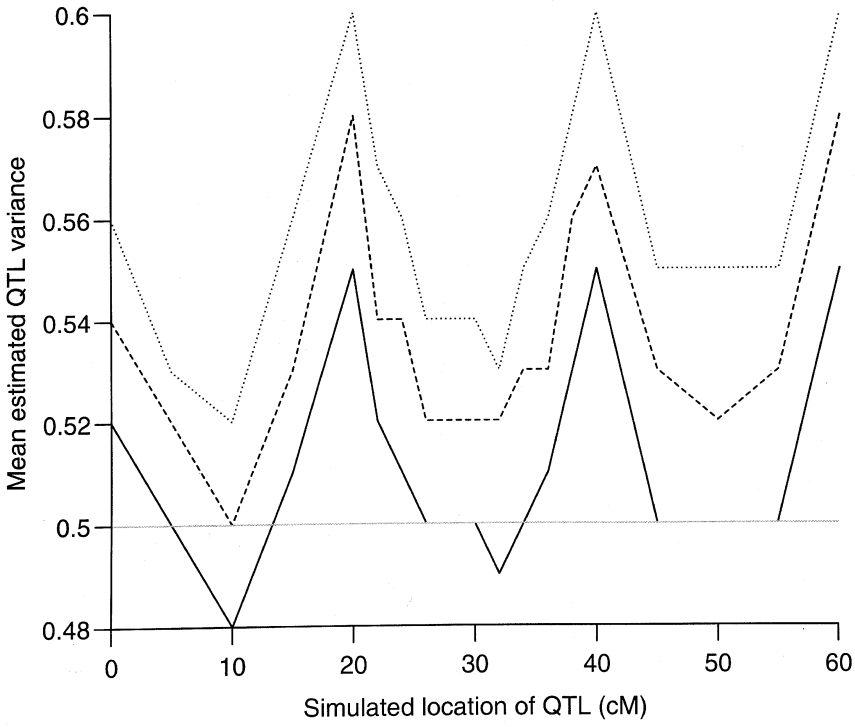

Fig. 3 The estimated QTL variance for the three methods for QTLs simulated at different locations in the linkage group. The results from the simple approach are shown by a solid line, the complete approach by a dotted line and the approach of Fulker *et al.* (1995) by a dashed line. The standard deviations of the variance estimates are around 0.16.

favourable attributes illustrated by Fulker *et al.* (1995). That is, they will give unbiassed estimates of location even when markers differ in their information content, and when the QTL is placed in an area of low information they may increase power.

In conclusion, we present a rapid and simple means of implementing sib-pair analysis for mapping of QTLs in outbred populations. Viewing the data in terms of identity by descent from the male and female parents separately allows extensions to the method to account for differences in recombination between sexes and also to accommodate both full- and half-sib data. This latter extension will be particularly important in domestic species, where a dataset may have large half-sibships and hence contain very large numbers of half-sib pairs. Extension to include half-sibs requires that both mean and variance differences in the distribution of sib-pair differences be accounted for, but the increase in power from inclusion of half-sibs can be appreciable (Hamann & Haley, 1998). The simplicity of this approach and its easy extension make it an obvious choice as a means of analysis with this type of data.

Acknowledgements

We are grateful for support from the Biotechnology and Biological Sciences Research Council, Ministry of Agriculture, Fish and Food, the European Commission and the Royal Society.

References

CHURCHILL, G. A. AND DOERGE, R. W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

FULKER, D. W. AND CARDON, L. R. 1994. A sib-pair approach to interval mapping of quantitative trait loci. *Am. J. Hum. Genet.*, **54**, 1092–1103.

FULKER, D. W., CHERNY, S. S. AND CARDON, L. R. 1995. Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am. J. Hum. Genet.*, **56**, 1224–1233.

HAMANN, H. AND HALEY, C. S. 1998. Combining full and half-sib data in sib pair linkage analysis. In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production* (Armidale, Australia). In press.

HASEMAN, J. K. AND ELSTON, R. C. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 3–19.