Theor Appl Genet (1995) 91:471–476

V. Hyne · M. J. Kearsey

QTL analysis: further uses of 'marker regression'

Received: 5 September 1994 / Accepted: 3 January 1995

Abstract A variety of approaches are available for identifying the location and effect of QTL in segregating populations using molecular markers. However, these have problems in distinguishing two linked QTL, particularly in relation to the size of the test statistic when many independent tests are performed. An empirical method for obtaining the distribution of the test statistic for specific datasets is described, and its power for demonstrating the inadequacy of a single-QTL model is explored through computer simulation. The method is an extension of the previously described technique of 'marker regression', and it is applied here to demonstrate two situations in which it may be useful. Firstly, we examine the power of the technique to distinguish two, linked QTL from one and compare this ability with that of two contemporary methods, 'Mapmaker/QTL' and 'regression mapping'. Secondly, we show how to combine information from two, or more, populations that may be segregating for different marker loci in a given linkage group. This is illustrated for two populations having in common just two linked marker loci although the sharing of loci is not a pre-requisite. Empirical tests are used to determine whether the same or different QTL are segregating and, if they are the same QTL, whether they are the same alleles. Evidence is discussed which suggests that the upper limit to the number of QTL that can be located for any single quantitative trait in a segregating populations is 12.

Key words Mapmaker/QTL · Marker regression · Numbers of QTL · QTL · Regression mapping

Communicated by G. Wenzel

V. Hyne (✉)
Horticulture Research International, Wellesbourne, Warwickshire CV35 9EF, UK

M. J. Kearsey
School of Biological Sciences, The University of Birmingham, Birmingham B15 2TT, UK

Introduction

There is growing interest in locating individual genes controlling quantitative traits (quantitative trait loci or QTL) given the availability of linkage maps comprising many molecular markers (Paterson et al. 1988; Hyne et al. 1994). A variety of statistical approaches have been described that favour using the information from multiple markers in a linkage group (Lander and Botstein 1989; Haley and Knott 1992; Martinez and Curnow 1992; Kearsey and Hyne 1994). These techniques yield comparable estimates of QTL position and effects (Haley and Knott 1992; Kearsey and Hyne 1994), although reliability is extremely poor unless very large populations are used (Hyne et al. 1995). However, the appropriate threshold value to apply in tests of significance is uncertain because numerous, often non-independent tests are performed. Without a suitable criterion for determining the presence of one QTL, it is impossible to test whether the effect detected is truly one QTL, or two (or more) linked QTL. In addition, a QTL detected in one cross cannot be tested for correspondence to a QTL found in a different cross. A method for establishing these threshold values is crucial to the interpretation of analytical results.

In a recent paper (Kearsey and Hyne 1994) we described marker regression, a simple approach to estimate QTL location and effects. We also mentioned that, using marker regression, a putative QTL could be tested for consistency with a one-QTL model and that estimates of QTL location and effects, detected in populations derived from different crosses, could be compared and distinguished. The present paper addresses these two issues by developing the theory and illustrating the method. The technique described applies to any population derived from an F_1, such as F_2, backcrosses, single-seed descent or doubled haploid (DH) lines. It attempts to overcome the problems of tests of significance by deriving empirical distributions of the test

statistic by computer simulation rather than modifying the significance levels.

Theory and methods

Consider an F_2 scored for a quantitative trait and a set of marker loci, k of which map to a particular linkage group. At each of these k loci, the mean trait value of individuals of the three genotypes $M_{i1}M_{i1}$, $M_{i1}M_{i2}$ and $M_{i2}M_{i2}$ can be calculated, where $i = 1, k$; 1 and 2 refer to the alleles from P_1 and P_2 respectively. As shown previously (Kearsey and Hyne 1994), the expected values of the additive effects (δ_i) and dominance deviations (λ_i) at a marker are,

$$\delta_i = \{M_{i1}M_{i1} - M_{i2}M_{i2}\}/2$$
$$= \{1 - 2R_i\}d$$
$$\lambda_i = \{M_{i1}M_{i2} - (M_{i1}M_{i1} + M_{i2}M_{i2})/2\}$$
$$= \{1 - 2R_i\}^2 h$$

where d and h are the additive and dominance effects of the QTL, and R_i is the recombination frequency between the ith marker and the QTL. If the positions of the QTL and the markers are known, regression of δ_i on to $(1 - 2R_i)$ should give a straight line of slope equal to d passing through the origin. Although the marker positions can be estimated, from either the marker data of that population or other sources, the QTL position is unknown. It can, however, be estimated by repeating the regression at regular intervals across the linkage group and identifying the position at which the residual mean square (RMS) is at a minimum. The residual MS for λ_i is very insensitive to QTL estimation, hence we have chosen to use δ_i in this work.

Extension to two, linked QTL

It is possible to test the adequacy of the model, based on one QTL on the chromosome. To illustrate this we need to extend the model to consider more than one QTL effect in a linkage group. If there are two QTL having a recombination frequency of R_{1i} and R_{2i} with marker i, than

$$\delta_i = (1 - 2R_{1i})d_1 + (1 - 2R_{2i})d_2$$

If the two QTL are tightly linked, then $R_{1i} = R_{2i} = R_i$ and hence

$$\delta_i = (1 - 2R_i)(d_1 + d_2)$$

and a plot of δ_i against marker position in centiMorgans (Haldane 1919) would yield the same curve as a single QTL with additive effect $(d_1 + d_2)$, as shown in Fig. 1a. However, if d_1 and d_2 are opposite in sign (i.e. QTL in dispersion) δ_i would approach zero at all marker positions.

More interestingly, if the two QTL are in association and further apart, the curve of δ_i will develop two peaks connected by a 'hanging valley', and the overall shape will fit that of the one QTL less well. This is shown in Fig. 1b for $d_1 = d_2 = 0.5$ at positions $C_1 = 30$ cM and $C_2 = 70$ cM. The problem is then to distinguish the curves in Fig. 1a and b. Unfortunately, unless the QTL

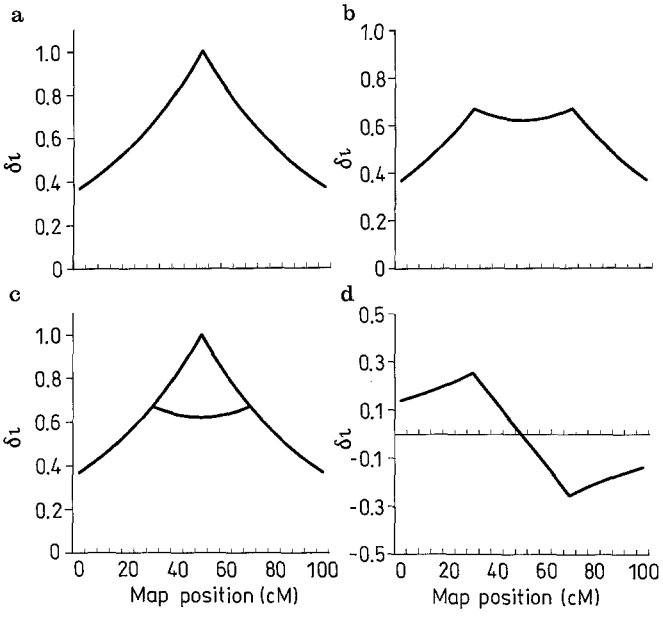

Fig. 1a–d Plot of marker differences δ_i against marker position for two QTL, $d_i = 0.5$, at various positions. **a** Two QTL, completely linked, in association at 50 cM, **b** two QTL at 30 cM and 70 cM in association, **c** a comparison of one QTL at 50 cM with two QTL at 30 cM and 70 cM, **d** two QTL at 30 cM and 70 cM in dispersion

effects at the two loci are very unequal, it is possible to replace the curves outside the two peaks by a curve produced from a single QTL of appropriate effect located between them, as shown in Fig. 1c. Therefore, discrimination between a model with one QTL and two QTL in association depends very critically on information from markers between the two QTL. If the two QTL are in dispersion, the curve resembles that in Fig. 1d and the problem is easier.

The data are first analysed using a single-QTL model in order to test whether the model is adequate and to identify the most likely QTL position and effect. At this position, a significant variance ratio for the residual MS will indicate that the one-QTL model does not fit the observed data. However, this variance ratio is a conservative test for two reasons, both of which would make the residual MS too small on average. Firstly, the QTL position is estimated where F is at a minimum and secondly, the Y (i.e. δ_i) values are not independent. It is difficult to see how to allow for these two factors theoretically but it is a simple matter to obtain the critical values of F empirically by computer simulation.

Suppose the initial analysis suggested a single QTL of effect $\hat{d}$ at position $\hat{C}$ cM with a variance ratio of F^* testing the residual MS. Using $\hat{d}$, $\hat{C}$ and the estimated marker positions, we can simulate a large number of F_2 populations of size equal to that of the actual population. Each set of simulated data is analysed identically to the observed data and the F values for the residual MS recorded. These will give the empirical F distribution, provided that the one-QTL model is correct, from which we can observe the probability of $F \geq F^*$.

Comparison of precision

In order to compare marker regression with 'interval mapping' for the ability to discriminate between two, linked QTL and a single QTL, data were simulated from a population of 300 individuals in which two QTL were embedded in a linkage group of six marker loci as shown below:

cM		36.70		84.55		
	|********|****[Q_1]****|********|****[Q_2]****|********|					
Marker	1	2	3	4	5	6
cM	0	25.54	47.85	73.39	95.70	121.24

Interval mapping was implemented using Mapmaker/QTL (Paterson et al. 1988) and 'regression mapping' (Haley and Knott 1992) fitting one and, where appropriate, two QTL models.

Power

The power of marker regression to identify two, linked QTL was examined using simulated F_2 populations of 300 individuals. Each individual had one chromosome 100 cM long and six marker loci each 20 cM apart. The distance between the two QTL was varied from 30 to 60 cM, and they were always separated by two markers. The QTL positions, effects and heritabilities are shown in Table 1. Ten replications of each situation were analysed.

Comparison of two or more populations

Consider two F_2 populations (A and B) of size n_A and n_B which have k_A and k_B markers, respectively, on a given chromosome. These markers may be all identical, all different or some, k_{AB}, may be common to both populations. First, each F_2 is analysed separately by marker regression. The data from the two populations are then combined to give $n_A + n_B$ individuals with $(k_A + k_B - k_{AB})$ marker loci and re-analysed, recording the variance ratio (F^*), QTL position (C_{A+B})and effect (d_{A+B}). Using these parameter estimates we can generate a large number of replicates (100 in this instance) of populations A and B of size n_A and n_B. The data from the 100 paired combinations of A and B are then analysed to obtain the empirical distribution of F in order to test the fit of the one-QTL, two-allele model.

Three situations were explored in which two populations were simulated to segregate for (1) the same alleles at the same locus; (2) different alleles at the same locus; and (3) different, linked loci. To achieve this, four different F_2 populations of 300 individuals were simulated to segregate at common markers 2 and 4 from the following set:

	|*****|******		|*****|*****|*****|*****|*****|*****|*****|*****|								
Marker	1	2	3	4	5	6	7	8	9	10	11
cM	0	10	20	30	40	50	60	70	80	90	100

Table 1 Marker loci, positions and effects of QTL in the four simulated F_2 populations (A, B, C, D)

Population	Marker										
	1	2	3	4	5	6	7	8	9	10	11
A	√	√	m	√	m	m *	√	√	√	m	m
B	m	√	m	√	√	√ *	m	m	m	√	√
C	m	√	m	√	√	√ †	m	m	m	√	√
D	m	√	m	√	√	√	m	m	m	√ *	√

m Monomorphic marker locus
√ Segregating marker locus
* QTL of additive effect 0.5 at this locus
† QTL of additive effect 1.0 at this locus

Population A segregated for markers 1, 2, 4, 7, 8 and 9 with one QTL, of additive effect 0.5 at marker 6. Populations B, C and D all segregated at markers 2, 4, 5, 6, 10 and 11. The QTL at marker 6 had an additive effect of 0.5 in population B (A + B = situation 1) and of 1.0 in population C (A + C = situation 2). Population D had a QTL of effect 0.5 segregating at marker 10 (A + D = situation 3). In all cases, the QTL accounted for 10% of the phenotypic variation and there was no dominance. Table 1 shows a summary of these population parameters. Each population was analysed independently and in combinations A + B, A + C and A + D. Populations A + C and A + D were further analysed, as described, to test the one-QTL, two-allele model.

Results

Comparison of precision

Figure 2 shows the results obtained after fitting one QTL to the simulated data set by the three methods. In all cases the curves had two peaks but one minimum (or maximum), at 58 cM for marker regression, at 64 cM for Mapmaker/QTL and at 62 cM for 'regression mapping'. When two QTL were fitted by regression mapping there was a significant improvement in fit ($P \leq 0.01$), although one of the QTL was located in the wrong interval. No attempt was made to fit two QTL using Mapmaker/QTL because the peak having the higher LOD score was incorrectly located (see Discussion). However, using marker regression, we were able to show that the one-QTL was inappropriate. To prove this, the method outlined above was applied as follows. From the initial marker regression analysis, the additive effect of the QTL was 0.8469, the environmental variance was 1.1368 and the variance ratio for the residual MS, $F^* = 2.055$. With these parameters, 100 replications of an F_2 population of 300 individuals were simulated. Of these, on only 2 occasions was $F \geq F^*$. Thus, if the model with one QTL were correct then an $F \geq F^*$ would occur on 2% of the occasions. Therefore, we can conclude that the single-QTL model, in this instance, does not provide an adequate fit to the data and that two, or more, linked QTL must be present.

Power

The probability with which two QTL were detected decreased as the distance between two QTL decreased

Fig. 2 A comparison of the three methods used to locate QTL when two QTL are segregating in a simulated dataset; LOD scores from Mapmaker/QTL, $pF_{\text{regression}}$ from 'regression mapping' and minimum residual MS from marker regression. Simulations are based on two QTL with additive effects of 0.5 at 36.7 cM and 84.5 cM in an F_2 population of 300 individuals. *i* Residual MS for marker regression, *ii dotted line* LOD score from Mapmaker/QTL, *solid line* $pF_{\text{regression}}$ from regression mapping.

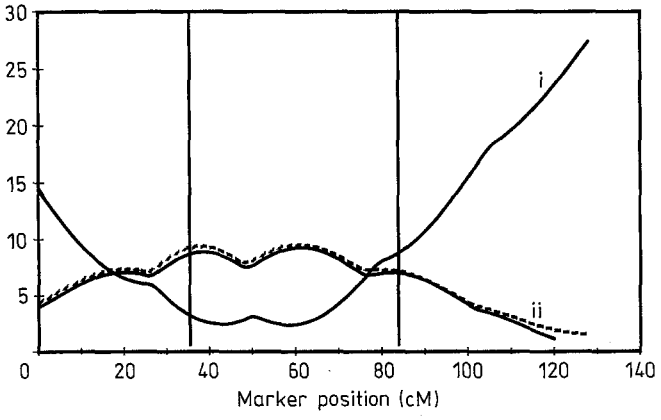

Table 2 Probability of success in distinguishing two QTL from one in F_2 populations of 300 individuals with different QTL positions and heritabilities

Distance apart (cM)	Q_1 (cM)	Q_2 (cM)	h_n^2 (%)	$P \leq 0.10$ (%)	$P \leq 0.05$ (%)	$P \leq 0.01$ (%)
50	25	75	10	100	100	90
40	30	70	10	80	60	50
30	35	65	10	20	10	0
60	20	80	5	100	100	60
50	25	75	5	50	40	30
40	30	70	5	50	40	0
30	35	65	5	10	0	0

(Table 2). When each QTL explained 10% of the phenotypic variation, the probability of resolving them was 100% at the 5% significance level when they were 50 cM apart. The probability of resolution dropped to 60% when the QTL were separated by 40 cM. When each QTL explained 5% of the phenotypic variation, the probability of resolving them was 100% when they were separated by 60 cM, reducing to 40% when 40 cM apart.

Comparison of two, or more populations

Table 3 shows the analysis of variance tables for populations A, B, C and D individually and in combinations of A + B, A + C and A + D. Analysis indicated that population A had a QTL of effect 0.4548 at 52 cM and population B had a QTL of effect 0.6083 at 48 cM. Both gave rise to an F^* value which was not significant, as expected, suggesting a good fit with the one-QTL model. Therefore, when combined, F^* was still not significant, leading to the correct conclusion that these two populations were segregating for the same allele for the trait. Two sources of evidence suggest that populations A and C had different alleles: d_C was more than twice the size of d_A, while F^* was highly significant (7.97) in the analysis of the combined populations. This can be seen graphically when δ_i is plotted against centiMorgan distance along the chromosome for A and C separately (Fig. 3a) and combined (Fig. 3b). When the two populations segregate for different loci (A and D) the regression on A and D separately indicate quite different positions for the QTL while a combined analysis yields an F^* of 2.14. This is illustrated in Fig. 3c and d. The empirical distribution of F taken from 100 simulations indicates that an $F^* \geq 2.14$ will occur on 13% of occasions and therefore has not achieved statistical significance in this case.

Discussion

We have previously shown how marker regression can be used to analyse datasets from experimental populations derived from an F_1 for the presence of QTL. Now,

Table 3 Marker regression analyses of variance for F_2 populations either individually or combined. See Table 1 and text for details

Population		*df*	MS	VR	Position	*d*
A	$d_{regression}$	1	70.2728	71.64	52.0	0.4548
	Residual	4	0.0435	0.04		
	Error	294	0.9810			
B	$d_{regression}$	1	147.9573	126.73	48.0	0.6083
	Residual	4	0.0288	0.02		
	Error	294	1.1675			
C	$d_{regression}$	1	597.7015	282.64	46.0	1.2170
	Residual	4	0.4043	0.19		
	Error	294	2.1147			
D	$d_{regression}$	1	64.2411	60.82	92.0	0.4536
	Residual	4	0.5809	0.55		
	Error	294	1.0562			
A + B	$d_{regression}$	1	373.2233	349.54	48.0	0.5492
	Residual	8	0.3913	0.37		
	Error	590	1.0678			
B + C	$d_{regression}$	1	972.1446	620.31	44.0	0.8976
	Residual	8	12.4979	7.97		
	Error	590	1.5672			
A + D	$d_{regression}$	1	227.3129	223.67	88.0	0.4576
	Residual	8	2.1766	2.14		
	Error	590	1.0163			

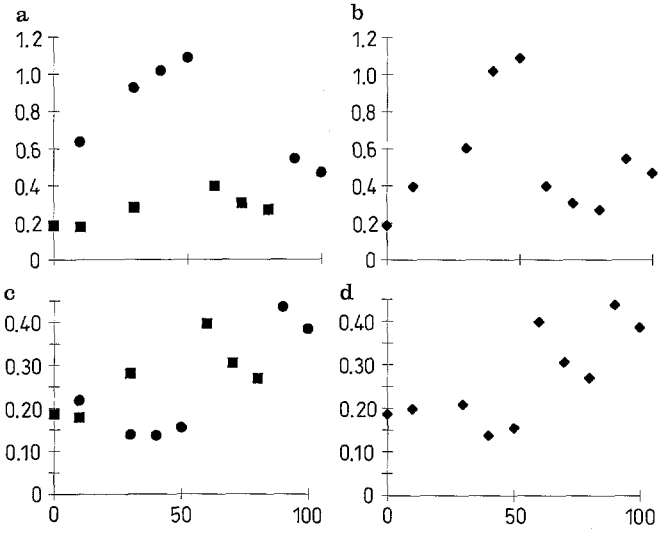

Fig. 3a–d Plots of δ_i against map position for two pairs of populations. Populations A and C had different alleles for the same QTL at 50 cM; A and D segregated for different QTL at 50 cM and 90 cM. **a** Populations A and C separately, **b** A and C combined, **c** A and D separately, **d** A and D combined. See text for details

in this paper, we demonstrate how the technique may be extended to test the fit of the one-QTL model for an effect detected in one linkage group and for effects detected in the same linkage group but in different populations. The facility to test the one-QTL model is available in Mapmaker/QTL, but there are drawbacks.

There are two possible interpretations for the detection of two peaks in one linkage group. The peaks could be linked QTL or the lower peak could be a local maximum of the higher peak (Fig. 2). Mapmaker/QTL provides a test which involves selecting "... one very likely QTL" and 'fixing' it in place (Mapmaker/QTL Version 1.1, Whitehead Institute (1993) whilst re-scanning the genome to detect other QTL. A LOD score greater than the sum of the LOD scores obtained by fitting each QTL individually supports the model of two, linked QTL acting independently. There is, however, no associated test of significance. In our example (Fig. 2), the 'very likely QTL' was selected as being that which had the higher LOD score. This was located in the wrong interval, hence to fix this QTL in position would be inappropriate. Regression mapping has no test for the adequacy of the one-QTL model, apart from the improvement in fit when two QTL are independently located. Marker regression, however, offers an empirical approach for testing the significance of the one-QTL model and is often correctly able to demonstrate the inadequacy of the one-QTL model when flanking marker methods fail.

Many factors affect the ability of a technique to distinguish between the presence of one and two QTL, the number of markers separating the QTL being one example. If no markers exist between them, two QTL will always be indistinguishable from one, irrespective of which analytical tool is employed. However, given that there are markers between the QTL, then the possibility exists of distinguishing the QTL and this increases as the heritability of the QTL and their distance apart increases. As Table 2 shows, the marker regression test of significance enables two, linked QTL, each responsible for 5% of the phenotypic variance, to be resolved at the 5% significance level on 100% of occasions when the QTL are 60 cM apart. This reduces to 40% when the

QTL are 40 cM apart while almost no resolution was possible when the QTL were separated by 30 cM. The 95% confidence interval for a single QTL of 5% heritability in a population of 300 F_2 individuals is approximately 60 cM (Hyne et al. 1995), hence the difficulty in separating two QTL just 30 cM apart is not surprising. Better resolution was predictably obtained with QTL each of 10% heritability. For example, at the 5% level, two QTL were distinguishable on 60% of the occasions when 40 cM apart.

Since marker regression is not dependent on flanking markers, information is not lost when genotyping is incomplete. Therefore, the approach would be applicable even if each marker locus in a linkage group were scored in a different set of F_2 individuals. Furthermore, it is possible that samples from an F_2 would be raised on different occasions and be scored for different markers as new techniques or probes become available. Provided that the trait does not exhibit excessive genotype × environment interaction the data can be combined. We have demonstrated that marker regression can be used to compare two or more populations derived from different crosses, grown in the same experiment, having as few as two markers in common. Allelic differences between populations were easily detected, and QTL 40 cM apart were detected on 87% of the occasions. Since different F_2 populations may well be polymorphic for different markers, such a simple comparison between populations is important. This is, indeed, a common occurrence in polyploid plants, such as hexaploid wheat, where different markers exist on homoeologous chromosomes and marker regression would be well-suited to the analysis of such situations. No approach to QTL location in an F_2 leads to very precise estimates of position or effect (van Ooijen 1992; Hyne et al. 1995), thus QTL located in different crosses may look to be very different even though they are really the same. The present method at least provides a way in which to check for consistency.

The varying power of methods to locate QTL with different heritabilities leads to an interesting conclusion regarding the number of QTL which one could hope to locate by any method. Consider a trait in an F_2 with an overall heritability of 0.4, which is not atypical of many traits, though possibly high for some. If we assume that this is due to the segregation of 8, 20 or 40 QTL of equal effect, then each QTL will have an individual heritability of 0.05, 0.02 or 0.01, respectively. Simulations have shown that approximately 95%, 60% or 30% of these QTL will be detected, i.e. 7, 12 or 12, respectively. It is unlikely, therefore, that more than 12 genes would ever be located for any one trait and normally, far fewer, because gene effects might be quite variable. This is supported by actual observations (Paterson et al. 1988; Hayes et al. 1993).

A set of programs and documentation for analysing F_2 and doubled haploid populations by marker regression is currently being prepared. We are happy to provide a copy of the executable versions upon request accompanied by a $3\frac{1}{2}''$ (90 mm) floppy disk.

References

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

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

Hayes PM, Liu BH, Knapp SJ, Chen F, Jones B, Blake T, Franckowiak J, Rasmusson D, Sorrells M, Ullrich SE, Wesenberg D, Kleinhofs A (1993) Quantitative trait locus effects and environmental interaction in a sample of Northern American barley germ plasm. Theor Appl Genet 87:392–401

Hyne V, Kearsey MJ, Martinez O, Gang W, Snape JW (1994) A partial genome assay for quantitative trait loci in wheat (*Triticum aestivum*) using different analytical techniques. Theor Appl Genet 89:735–741

Hyne V, Kearsey MJ, Pike DJ, Snape JW (1995) QTL analysis: unreliability and bias in estimation procedures. Plant Molec Biol (in press)

Kearsey MJ, Hyne V (1994) QTL analysis: A simple marker regression approach. Theor Appl Genet 89:698–702

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

Martinez O, Curnow RNC (1993) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480–488

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721–726

van Ooijen JW (1992) Accuracy of mapping quantitative trait loci in autogamous species. Theor Appl Genet 84:803–811

Whitehead Institute (1993) Mapping genes controlling quantitative traits using Mapmaker/QTLversion 1.1: a tutorial and reference manual. Whitehead Institute, Cambridge, Mass