Theor Appl Genet (1994) 89: 698–702

M. J. Kearsey · V. Hyne

QTL analysis: a simple 'marker-regression' approach

Received: 10 February 1994 / Accepted: 17 May 1994

Abstract A method to locate quantitative trait loci (QTL) on a chromosome and to estimate their additive and dominance effects is described. It applies to generations derived from an F_1 by selfing or backcrossing and to doubled haploid lines, given that marker genotype information (RFLP, RAPD, etc.) and quantitative trait data are available. The method involves regressing the additive difference between marker genotype means at a locus against a function of the recombination frequency between that locus and a putative QTL. A QTL is located, as by other regression methods, at that point where the residual mean square is minimised. The estimates of location and gene effects are consistent and as reliable as conventional flanking-marker methods. Further applications include the ability to test for the presence of two, or more, linked QTL and to compare different crosses for the presence of common QTL. Furthermore, the technique is straightforward and may be programmed using standard pc-based statistical software.

Key words Molecular markers · QTL location
Quantitative traits · Regression mapping

Introduction

The availability of abundant, naturally-occurring, molecular genetic markers (RFLPs, RAPDs, isozymes, etc) during the last decade has generated renewed activity into counting, locating and measuring the effects of genes (polygenes or QTL) controlling quantitative traits (Beckmann and Soller 1986; Weller 1986; Edwards et al. 1987; Paterson et al. 1991; Simpson 1989; Luo and Kearsey 1991; Haley and Knott 1992; Martìnez and Curnow 1992; Stuber et al. 1987). Interest particularly surrounds traits of economic importance in crop plants and domesticated animals.

Currently, the most popular analytical method to investigate QTL is that of flanking-marker mapping, either by the log-likelihood approach of interval mapping, as implemented by Mapmaker/QTL (Lander and Botstein 1989), or by multiple regression (Haley and Knott 1992; Martìnez and Curnow 1992). The two methods yield very similar results (Haley and Knott 1992; Martìnez 1994), but the multiple-regression approach applies a more straightforward test of significance and is programmable using standard statistical packages.

The efficiency of flanking-marker methods decreases as the number of incompletely-genotyped individuals increases. This difficulty has, in part, been overcome by a technique developed by Martìnez and Curnow (1994). Nonetheless, such approaches have problems separating one QTL from two on the same chromosome, and the use of three, or more, marker regression is advised (Martìnez and Curnow 1992). The latter has the drawback of requiring large population sizes because the expected probability of certain genotypic classes occurring can be very small. Furthermore, these methods are not capable of combining the information from two, or more, populations, each having different markers.

The present paper describes a 'marker-regression' approach which can be used for populations derived from an F_1. It is as reliable as the interval-mapping and multiple-regression approaches, but has wider application and is capable of hypothesis testing. It also relies on simple statistical procedures, using standard software.

Theory

Consider a pair of homologous chromosomes in an F_1 produced from a cross between two, true-breeding lines, P_1

Communicated by J. W. Snape

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

V. Hyne
Horticulture Research International,
Wellesbourne,
Warwick CV35 9EF, UK

and P_2. Let this pair of chromosomes be heterozygous for alleles at k marker loci, M_{i1} and M_{i2}, depending on whether the allele came from P_1 or P_2 respectively and $i=1,k$, situated at C_i cM (Haldane) on the linkage map of that chromosome. Finally, let there be a single QTL (Q_1Q_2) on this chromosome at X cM.

An F_2 of N individuals is derived from this F_1 and every individual is scored for a quantitative trait, Y_j (where $j=1$ to N), and its marker genotype determined at each of the k marker loci. From the latter, the map positions of the markers (C_i) can be estimated (e.g. by Mapmaker; Lander et al. 1987).

Let the mean trait score of the three possible QTL genotypes in the F_2 be as follows:

$$Q_1Q_1 = \mu + d$$
$$Q_2Q_2 = \mu - d$$
$$Q_1Q_2 = \mu + h$$

where μ is the mean of the two homozygotes and d and h are the additive and dominance effects respectively, as defined by Mather and Jinks (1982) except that, since either Q_1 or Q_2 may have the larger effect, d and h may be either positive or negative.

Because it is not possible to genotype the QTL we have to rely on marker phenotypes, and the present procedure uses the mean scores of each of the three genotypes at every marker locus. Following standard theory (e.g., Cowen 1988)

$$\overline{M_{i1}M_{i1}} = \mu + (1 - 2R_i)\, d + 2R_i(1 - R_i)\, h$$
$$\overline{M_{i2}M_{i2}} = \mu - (1 - 2R_i)\, d + 2R_i(1 - R_i)\, h$$
$$\overline{M_{i1}M_{i2}} = \mu + [1 - 2R_i\,(1 - R_i)]\, h$$

where $\overline{M_{i1}M_{i1}}$, etc., is the expected mean trait value of all those individuals having marker genotype $M_{i1}M_{i1}$, where $i=1$ to k, and R_i is the recombination frequency between the QTL and the ith marker. Therefore,

$$\tfrac{1}{2}\,(\overline{M_{i1}M_{i1}} - \overline{M_{i2}M_{i2}}) = (1-2R_i)\,d = \delta_i \qquad (1)$$

$$\overline{M_{i1}M_{i2}} - \tfrac{1}{2}\,(\overline{M_{i1}M_{i1}} + \overline{M_{i2}M_{i2}}) = (1-2R_i)^2\, h = \lambda_i \qquad (2)$$

Following Haldane (1919), $(1-2R_i)=e^{-m}$, where $m=|(X-C_i)/50|$, the mean chiasma frequency in that interval.

The relationship between δ_i, λ_i and marker position in cM is shown by the curve in Fig. 1(a) for a QTL at 50 cM with gene effects $d=h=1$. In practice, few markers are present and the observed outcome for a possible set of six markers is illustrated by the bars of δ_i in Fig. 1(a).

The present approach, to detect, locate and estimate the effect of a QTL, is based on finding the values of X, d and h which best fit the observed values of δ_i and λ_i at map positions C_i.

We note from equation (1) that $\delta_i=(1-2R_i)d$. Thus, if $R_i=0.5$, i.e., there is no linkage between the QTL and the ith marker, then $\delta_i=0$; if $R_i=0$, i.e., complete linkage between the QTL and the marker, $\delta_i=d$. Equation (1) is thus a linear equation of the form $y=0+bx$. Therefore, if we regress δ_i, i.e., y, on $(1-2R_i)$, i.e., x, we should obtain a straight line of slope $b=d$ passing through the origin. Similarly, from equation (2), $\lambda_i=(1-2R_i)^2h$. This is also a linear equation $y=0+bx$ where $y=\lambda_i$ and $x=(1-2R_i)^2$. Therefore, regression yields $b=h$. It should be noted that λ_i and δ_i are orthogonal and are, hence, independent variables.

In such a situation, the uncorrected sums of squares and products should be used to perform the regression analysis, i.e.,

$$\hat{b} = \Sigma xy/\Sigma x^2 = \Sigma\,\delta_i\,(1-2R_i)/\Sigma\,(1-2R_i)^2,$$

$$\text{Regression SS} = \hat{b}\Sigma xy = [\Sigma\,\delta_i\,(1-2R_i)]^2/\Sigma\,(1-2R_i)^2,$$

$$\text{Residual SS} = \Sigma y^2\text{-Regression SS}$$
$$= \Sigma\,\delta_i^2 - [\Sigma\,\delta_i\,(1-2R_i)]^2/\Sigma\,(1-2R_i)^2.$$

The degrees of freedom (df) for the regression and remainder are 1 and $k-1$ respectively but we will need to reconsider them later.

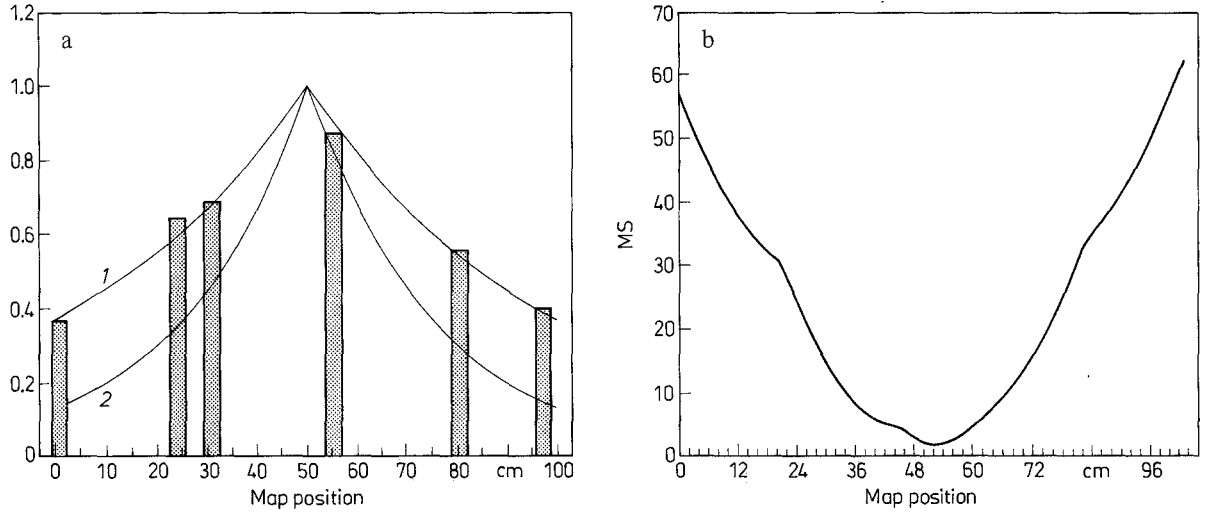

Fig. 1 a, b The effects of one QTL on δ_i, λ_i and residual mean squares associated with genetic markers. (**a**) Relation between δ_i, λ_i and marker position (*curves 1 and 2*, respectively). *Bars* indicate possible effects associated with marker loci. (**b**) Change in residual mean square for various putative QTL positions

However, since the true position, X, of the QTL is unknown, the true values of R_i can not be calculated. Nonetheless, the regression analysis of variance for δ_i may be carried out at a range of possible positions of the QTL along the chromosome from which R_i can be calculated. The estimated QTL location is then the position at which the residual SS is at a minimum.

Example

The procedure, as outlined above, is illustrated for a set of simulated data. The F_2 population generated had a single QTL situated at 50 cM from the left-most marker locus, having d=1.0, h=0.5 and a narrow-sense heritability (h_n^2) of 0.1, i.e., the expected phenotypic variance of the F_2 was 5.0. Six, equally-spaced markers (k=6) were set at 0, 20, 40, 60, 80 and 100 cM. The data in Table 1 are based on a random sample of 300 F_2 individuals and give the individual marker-genotype means for the trait at each locus, together with the estimated δ_i values. The cumulative marker positions along the chromosome, estimated from the data, are shown in Table 2. Any differences between these estimated positions and those set above are due to sampling but the estimated values are used in subsequent computations to mimic a real situation.

Table 2(a) illustrates the data for regressing δ_i on $(1-2R_i)$ for a putative QTL at 30 cM. Using the sums of squares and products in Table 2(b), the regression analysis of variance in Table 3(a) was obtained. Since δ_i was calculated from the phenotypic means of two genotypic classes, the values obtained for the mean squares must be adjusted in line with the error, V_E, which was based on individuals. For example, since

$$\delta_i = \tfrac{1}{2}(\overline{M_1M_1} - \overline{M_2M_2}),$$

then the variance of δ_i, $V_{\delta i} = \tfrac{1}{4}(V_{\overline{M_1M_1}} + V_{\overline{M_2M_2}})$.
Now

$$V_{\overline{MiMj}} = V_E/n$$

where n is the number of individuals of that genotype in the F_2 and is expected to be $\frac{1}{4}N$ for M_1M_1 or M_2M_2 and $\frac{1}{2}N$ for M_1M_2.

Thus,

$$V_{\delta i} = \tfrac{1}{4}[V_E/\tfrac{1}{4}N + V_E/\tfrac{1}{4}N] = 2V_E/N.$$

Similarly,

$$V_{\lambda i} = 4V_E/N.$$

Consequently, the SS in the regression ANOVA must be adjusted by multiplying by $\frac{1}{2}N$ (i.e., 150) for δ_i and by $\frac{1}{4}N$ (i.e., 75) for λ_i, since the F_2 was of size N=300.

The error mean square, V_E, is simply the F_2 variance less the genetical variance at the QTL. The genetical variance for a single locus is $1/2d^2 + 1/4h^2$ and, in the results of Table 3(a), $\hat{d} = 1.3679$, $\hat{h} = 0.0122$ and $\hat{V}_{F_2} = 4.9677$; thus, $\hat{V}_E = 4.03$.

We see from Table 3(a) that the 'residual' MS for δ_i is highly significant, indicating that fitting a model involving a single QTL at that position, 30 cM, does not explain the δ_i and λ_i values observed, even though the estimate of d is highly significant.

A computer program may be written to repeat this analysis for a putative QTL at regular positions along the chromosome and compute the residual MS at each position. The graph of residual MS for δ_i against position on the chromosome is shown in Fig. 1(b). There is a minimum at 54 cM indicating the most-likely position of the QTL, and the

Table 1 Marker genotype means

Marker locus	Marker genotype			δ_i	λ_i
	M_1M_1	M_1M_2	M_2M_2		
1	23.5669	22.8834	22.5475	0.5097	–0.1738
2	23.9616	22.8657	22.3639	0.7988	–0.2970
3	23.8994	23.0770	21.6758	1.1118	0.2894
4	24.0287	23.0861	21.6767	1.1760	0.2334
5	23.6174	23.0387	22.1215	0.7480	0.1692
6	23.1662	23.0273	22.5919	0.2871	0.1482

Table 2 Illustration of linear regression approach. (a) Data set with six marker loci and a putative QTL at X=30 cM and (b) Regression SS and SP. Data for δ_i and λ_i from Table 1

(a)

Marker locus m (i)	Marker Position (c_i)	Distance of QTL from m (i) $\lvert(x-c_i)\rvert$	(x_d) $(1-2R_i)$	(y_d) δ_i	(x_h) $(1-2R_i)^2$	(y_h) λ_i
1	0.0	30.0	0.5488	0.5097	0.3012	– 0.1738
2	20.6	9.4	0.8286	0.7988	0.6866	– 0.2970
3	45.2	15.2	0.7379	1.1118	0.5445	0.2894
4	64.5	34.5	0.5016	1.1760	0.2516	0.2334
5	82.5	52.5	0.3499	0.7480	0.1224	0.1692
6	103.7	73.7	0.2290	0.2871	0.0524	0.1482

(b)

	δ_i	λ_i
$\sum x^2$	1.9587	0.9397
$\sum y^2$	4.1589	0.3072
$\sum xy$	2.6794	– 0.0115

Table 3 Regression analyses for δ_i and λ_i. (a) for a QTL at X=30 cM and (b) for a QTL at X=54 cM (the optimum). V_{F2}=4.9677 for 299 *df*

(a) ANOVA

Source	*df*	SS	MS	Adj MS†	*F*	Sign
δ_i Regression	1	3.6653	3.6653	549.80	37.1	***
δ_i Residual	5	0.4936	0.0987	14.81	3.7	**
λ_i Regression	1	0.0001	0.0001	0.01	< 0.1	ns
λ_i Residual	5	0.3071	0.0614	4.61	1.1	ns
Error	287			4.03		

Parameter estimates: d=1.3679; h=0.0122
† For δ_i, MS × ½N where N=size of F_2 (300)
For λ_i, MS × ¼N where N=size of F_2 (300)

(b) ANOVA

Source	*df*	Adj MS	*F*	Sign
δ_i Regression	2	306.9805	76.50	***
δ_i Residual	4	2.4736	0.62	ns
λ_i Regression	1	7.3693	1.84	ns
λ_i Residual	5	3.1344	0.78	ns
Error	287	4.013		

Parameter estimates: d=1.3654; h=0.2949
** $P \leq 0.01$; *** $P \leq 0.001$; ns=not significant

Table 4 Estimates of QTL position and effects from marker regression and Mapmaker/QTL based on 100 simulations of two genetic models, together with the correlations between methods (r:98 *df*)

h_n^2	Expected		Estimates (standard deviation)				r
			Marker regression		Mapmaker/QTL		
0.1	*X*	50.0	50.16	(8.20)	49.80	(7.80)	0.68
	d	1.0	1.00	(0.20)	1.01	(0.20)	0.79
	h	0.5	0.46	(0.33)	0.49	(0.35)	0.75
0.05	*X*	62.2	63.20	(13.02)	63.10	(10.83)	0.54
	d	0.5	0.52	(0.16)	0.52	(0.14)	0.89
	h	0.5	0.46	(0.23)	0.50	(0.23)	0.83

analysis of variance at this point is shown in Table 3(b). However, because we have estimated the map position of the QTL, X, the regression SS now has 2*df*, one each for the estimation of b and X, while the residual SS has k–2*df*, as explained by Haley and Knott (1992). The residual MS is not significant for either δ_i or λ_i while the additive effect of the QTL, d, tested by the regression, is highly significant and positive, indicating that Q_1 is the increasing allele. The dominance effect is also positive but not significant. The data are thus consistent with a QTL at 54 cM with $\hat{d}$=1.3654 and $\hat{h}$=0.2949 (although the latter is not significantly different from zero).

Although a similar analysis was performed to find the point of minimum residual MS for λ_i, a very flat curve with an imprecisely-defined minimum was obtained. The estimates of dominance effects were therefore based on the X value determined from δ_i.

Test of reliability

The reliability of the marker-regression approach was tested by computer simulation. A wide range of genetic situations was examined, varying in the location of the QTL, its heritability and genetic effects. One hundred replicate samples, each of 300 F_2 individuals, were simulated from every genetic situation. The QTL location and the gene effects of each sample were estimated using both the present marker-regression method and Mapmaker/QTL, and the mean and standard deviation of the estimates over replicates, together with the correlations between the two methods, were computed. The results for two genetic situations, with contrasting parameters, are illustrated in Table 4, but they were typical of all models explored.

Table 4 shows that the estimates of QTL position and genetic effects, obtained by the present approach, are consistent and are comparable to those of Mapmaker/QTL, both in mean and standard deviation. Moreover, the estimates from the two methods are significantly correlated over simulations. The correlation coefficient is far from unity, an expected result since different sub-sets of the data were used. By inference from the results of Haley and Knott (1992) and Martìnez and Curnow (1992), therefore, the results of the present technique should be comparable to those obtained by flanking-marker regression.

Tests of significance

As discussed by others (Lander and Botstein 1989; Haley and Knott 1992), the significance level to be used in these

types of analysis is not yet clear. In fact the latter authors simply state that, " ... large values of (the test statistic) support the hypothesis that a QTL is present".

Because the number of independent tests is high, it is conventional to set the significance level at 0.001, i.e., adequate for at least 50 tests with a combined probability of 0.05. In the present context, with a regression SS having $2df$ and an error variance with a very large number of df, this requires $\chi^2[2]$ of 13.8 (or F of 6.9). Simulations with no QTL segregating on the chromosome indicate that this is a reasonable criterion as it leads to only 3% false positives. Similar simulations of a single QTL with individual heritabilities of 0.05, 0.02 and 0.01 result in a significant QTL being identified on approximately 95%, 60%,and 30% of occasions, respectively. Such power is at least as good as other methods.

Discussion

Since this marker-regression approach produces estimates of QTL location and effects which are comparable to existing methods, what are its advantages? The approach is, it is hoped, simple to understand and easy to program using standard statistical software. Since it does not depend on flanking-markers, it offers several unique features, some of which are outlined below.

First, the residual mean square can be used to test the adequacy of the simple one-QTL model on a given chromosome. It incorporates all the marker information on that chromosome in a single test. Second, it provides a simple test for whether the QTL, located on a given chromosome in different populations, are the same and this is achieved by standard joint regression analysis. Different populations typically segregate at different marker loci and hence create difficulties for flanking-marker methods. It is of no consequence to marker regression whether the markers are in common or are completely different. Both these features will be developed and illustrated in a subsequent paper.

Because we are using samples from the same data set to estimate the marker means, the error variances are correlated and hence the normal rules for analysis of variance are violated. However, with data sets of the size used here, the effects are small and result in conservative tests of significance of the one-QTL model. These correlations do not bias QTL location. A method to improve the power of the test for more than one linked QTL will be considered in a subsequent paper.

We have illustrated the procedure for an F_2 population, but the approach is easily adapted for other generations derived from an F_1, e.g., backcrosses, doubled haploid lines or single-seed descent lines. In every case a linear model can be written of gene effect against a function of recombination frequency between the QTL and the marker, as follows (See Cowen 1988):

(i) backcross to P_1: (3)
$$\tfrac{1}{2}\,(\overline{M_{i1}M_{i1}} - \overline{M_{i1}M_{i2}}) = \tfrac{1}{2}\,(1-2R)\,(d - h)$$
to P_2: (4)
$$1/2\,(\overline{M_{i1}M_{i2}} - \overline{M_{i2}M_{i2}}) = \tfrac{1}{2}\,(1-2R)\,(d + h)$$
$$(3) - (4) = (1-2R)\,d$$

(ii) DH Lines: $\tfrac{1}{2}\,(\overline{M_{i1}M_{i1}} - \overline{M_{i2}M_{i2}}) = (1-2R)\,d$ (5)

(iii) SSD Lines: $\tfrac{1}{2}\,(\overline{M_{i1}M_{i1}} - \overline{M_{i2}M_{i2}}) = [(1-2R)/(1+2R)]\,d.$ (6)

Finally, although we have confined attention to QTL and markers on one chromosome, the method can be applied sequentially to all chromosomes once the linkage groups have been ascertained. The error variance, V_E, used earlier, will include variation from genetic segregation at all other unlinked QTL. However, when several QTL have been located by this approach their combined effects can be removed to minimise V_E and so improve efficiency.

References

Beckmann JS, Soller M (1986) Restriction fragment length polymorphisms in plant genetic improvement. Oxford Surv Plant Mol Cell Biol 3:197–246

Cowen NM (1988) The use of replicated progenies in marker-based mapping of QTLs. Theor Appl Genet 75:857–862

Edwards MD, Stuber CW, Wendel JF (1987) Molecular-marker-facilitated investigation of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125

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

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

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER; an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Luo ZW, Kearsey MJ (1991) Maximum-likelihood estimation of linkage between a marker gene and a quantitative trait locus. II. Application to backcross and doubled haploid populations. Heredity 66:117–124

Martìnez O (1994) Quantitative trait loci estimation in plant populations. PhD thesis, University of Reading, UK

Martìnez 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

Martìnez O, Curnow RN (1994) Missing marker data when estimating quantitative trait loci using regression mapping. Heredity 73:198–206

Mather K, Jinks JL (1982) Biometrical genetics. Chapman and Hall, London

Paterson AH, Damon S, Hewitt JD, Zamir D, Rabinowitch HD, Lincoln SE, Lander ES, Tanksley SD (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics 127:181–197

Simpson SP (1989) Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. Theor Appl Genet 77:815–819

Stuber CW, Edwards MD, Wendel JF (1987) Molecular-marker-facilitated investigation of quantitative trait loci in maize. II. Factors influencing yield and its component traits. Crop Sci 27:639–648

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