W. R. Wu · W. M. Li

A new approach for mapping quantitative trait loci using complete genetic marker linkage maps

Received: 21 January 1993 / Accepted: 26 October 1993

Abstract A new approach based on nonlinear regression for the mapping of quantitative trait loci (QTLs) using complete genetic marker linkage maps is advanced in this paper. We call the approach joint mapping as it makes comprehensive use of the information from every marker locus on a chromosome. With this approach, both the detection of the existence of QTLs and the estimation of their positions, with corresponding confidence intervals, and effects can be realized simultaneously. This approach is widely applicable because only moments are used. It is simple and can save considerable computer time. It is especially useful when there are multiple QTLs and/or interactions between them on a chromosome.

Key words Quantitative trait locus (QTL) · Genetic marker · Joint mapping · Genetic linkage map

Introduction

Systematic and efficient mapping of quantitative trait loci (QTLs) depends on the establishment of complete genetic marker linkage maps. Research on QTL mapping has progressed slowly in the past because of the lack of such complete maps. Since the 1980s, however, rapid progress has have been made in this field due to the discovery of restriction fragment length polymorphism (RFLP) – a kind of highly-variable molecular genetic marker, which enables complete genetic marker linkage maps to be established (Beckman and Soller 1986; Tanksley et al. 1989). To-date, highly-saturated RFLP linkage maps with wide coverage of chromosomes have been obtained in many plants. It is anticipated that complete linkage maps with a high density of marker loci will be established for most of the important crops in the near future.

QTL mapping involves detecting the existence of QTLs and estimating their exact positions. A number of different approaches have been proposed both for QTL detection, such as the t-test (Ellis 1986; Simpson 1989) and ANOVA (Li et al. 1993), and for QTL location, such as maximum likelihood estimation (Weller 1986; Luo and Kearsey 1989), moment estimation (Mather and Jinks 1971; Snape et al. 1985; Wu et al. 1991), and regression analysis (Haley and Knott 1992). Moreover, the likelihood ratio test (Simpson 1989; Lander and Botstein 1989) can be used for both the detection and the location of QTLs.

Most of the methods which have been developed are based on the model of one marker locus linked to one QTL. Since they exploit the statistical information of each marker separately they are less efficient and less reliable (Lander and Botstein 1989). They may even give false results when there are more than one QTL on the same chromosome contributing to the genetic variation. These approaches, therefore, are not ideal given that only nearly complete marker linkage maps have been availabe.

For this reason, Lander and Botstein (1989) advanced an approach, called interval mapping, which could overcome the drawbacks mentioned above. Some studies have been made with this approach (Luo and Kearsey 1992; Carbonell et al. 1992), but the approach itself is not perfect. First, like other approaches based on maximum likelihood estimation, the hypothesis of the normality of the phenotypic distribution of a QTL genotype in the model may not always be true because there is often dominance and epistasis among polygenes. Second, the statistical meaning of a one-lod support interval is not clear enough because it does not indicate the probability level of confidence. Third, the determination of a suitable threshold for hypothesis testing is

Communicated by A. L. Kahler

W. R. Wu[1] (✉) · W. M. Li
Department of Agronomy, Fujian Agricultural College, Fuzhou, Peoples Republic of China

complex and difficult. And lastly, its computation is both complicated and time-consuming. Recently, Haley and Knott (1992) have proposed a new method of interval mapping based on regression, which makes the computation much simpler without, apparently, losing the efficiency of estimation. But it still faces the problem of threshold determination.

In this paper we advance a new approach to QTL mapping, in which both detection and location can be realized simultaneously. Because it makes comprehensive use of the information provided by all the markers on the same chromosome, it is named the joint mapping approach.

Principles

Material

In this paper, only the first generation backcross population (BC_1, or $F_1 \times P_1$) will be discussed. In such a population, two alleles (A_1 and A_2) of a locus segregate into two genotypes with equal proportions (1 A_1A_1:1 A_1A_2). A QTL is considered to exist only when it consists of two different alleles so that it contributes to the variation of the related quantitative trait.

One-QTL-mapping

Mathematical model Assume that in BC_1 the two genotypes of a QTL (Q_1Q_1 and Q_1Q_2) have means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. For a given marker locus, the means and variances of its two genotypes (G_1G_1 and G_1G_2) can be written as follows:

$$M_1 = (1-r)\mu_1 + \mathrm{r}\mu_2 \tag{1}$$

$$M_2 = r\mu_1 + (1-r)\mu_2 \tag{2}$$

$$V_1 = (1-r)\sigma_1^2 + r\sigma_2^2 + r(1-r)(\mu_1-\mu_2)^2 \tag{3}$$

$$V_2 = r\sigma_1^2 + (1-r)\sigma_2^2 + r(1-r)(\mu_1-\mu_2)^2 \tag{4}$$

where r denotes the recombination rate between the marker and the QTL. Theoretically, it is often assumed that $\sigma_1^2 = \sigma_2^2 = \sigma^2$; hence, from (3) and (4), we have

$$V_1 = V_2 = \sigma^2 + r(1-r)(\mu_1-\mu_2)^2. \tag{3}$$

From (1) and (2) it can be found that

$$M_1 - M_2 = (1-2r)(\mu_1-\mu_2). \tag{6}$$

Let $y = M_1 - M_2$, $x = 1-2r$, and $b = \mu_1 - \mu_2$ (we define it as the effect of the QTL), then (6) becomes

$$y = bx. \tag{7}$$

Assuming Haldane's map function,

$$r = \tfrac{1}{2}[1 - \exp(-0.02|z - z_Q|)] \tag{8}$$

where z and z_Q are the position coordinates on the map (unit: cM) of the marker and the QTL, respectively. Thus, in (7)

$$x = \exp(-0.02|z - z_Q|). \tag{9}$$

So y is determined by b, z and z_Q.

Since the position of each marker is already known on a complete linkage map, a set of parired data (y_i, z_i), $i = 1, 2, \ldots, n$ (assuming a total of n markers), can be obtained in an experiment. Therefore, according to (7) and (9), a two-parameter nonlinear regression model can constructed as follows:

$$y_i = bx_i(z_Q) + \varepsilon_i \tag{10}$$

where ε_i is the residual error. Since y_i is the difference between two means, it should show asymptotically a normal distribution, namely $y_i \sim N[bx_i(z_Q), \sigma_{\varepsilon i}^2]$, or $\varepsilon_i \sim N[0, \sigma_{\varepsilon i}^2)$, in accordance with the central limit theorem. From (5) it is possible to show that

$$\sigma_{\varepsilon i}^2 = 2[\sigma^2 + r_i(1-r_i)(\mu_1-\mu_2)^2]/(N/2) \tag{11}$$

where N is the sample size. The formula shows that $\sigma_{\varepsilon i}^2$ is related to r_i so that it varies with different marker loci.

Estimation of parameters Since $\sigma_{\varepsilon i}^2$ is not equal to each other, the approach of weighted least square (WLS) is the best way to estimate the parameters b and z_Q in model (10). The residual sum of square is

$$Q = \sum_{i=1}^{n} \varepsilon_i'^2 = \sum_{i=1}^{n} w_i[y_i - bx_i(z_Q)]^2 \tag{12}$$

in which $\varepsilon_2' = \varepsilon_i/\sigma_{\varepsilon i}$, $w_i = 1/\sigma_{\varepsilon i}^2$. Obviously, $\varepsilon_i' \sim N(0, 1)$. Hence ε_i' is an *iid*$N(0, 1)$.

For a given value of z_Q, the WLS estimate of b is

$$\hat{b} = \frac{\sum w_i x_i(z_Q) y_i}{\sum w_i [x_i(z_Q)]^2}. \tag{13}$$

Changing the value of z_Q with a step length of 1 cm along the whole linkage map, the point $\hat{z}_Q$ which makes Q minimum can be found; that is, $Q_{\min} = Q(\hat{b}, \hat{z}_Q)$.

Test of significance Since ε_i' is an *iid*$N(0, 1)$, the $Q_{\min}$ must be distributed as a χ^2 $(n-2)$ (note that two degrees of freedom are lost here because two parameters have been estimated). The fitness of the model, therefore, can be tested by mens of a χ^2-test; that is, acceptance of the model requires $Q_{\min} < \chi_{0.05}^2(n-2)$.

If the model is fit, a further test of significance of $\hat{b}$ is needed. Here the null hypothesis is H_0: $b = 0$ (there is no QTL), and the alternative hypothesis is H_1: $b \neq 0$ (there is a QTL). Similarly, since ε_i' is an *iid*$N(0, 1)$, there would be approxamitely (Johansen 1984)

$$Q(b=0) - Q_{\min} \sim \chi^2(1) \tag{14}$$

where $Q(b=0)$ denotes the minimum residual sum of squares when $b = 0$, which, refering to (12), equals to $\Sigma w_i y_i^2$. Thus, when $Q(b=0) - Q_{\min} \geq \chi_\alpha^2(1)$, in which α is a given significance level, we will refuse H_0 but accept H_1.

Determination of the significance level depends on the number of chromosomes involved. A higher level is required when the number increases, because considering many chromosomes at the same time may increase the risk that false positives will occur. The overall null hypothesis is that there are no QTLs on all chromosomes. If the required overall significance level is α_0, then the nominal significance level must be

$$\alpha = 1 - (1 - \alpha_0)^{1/m} \tag{15}$$

where m is the number of the chromosomes.

Determination of confidence intervals Desides the point estimate of a QTL's position ($\hat{z}_Q$), we also need to know its confidence interval on the chromosome. Since ε_i' is an *iid*$N(0, 1)$, it follows (Hamilton 1986) that $Q(z_Q) - Q_{\min} \sim \chi^2(1)$, where $Q(z_Q)$ denotes the minimum residual sum of squares depending on z_Q. And hence the 95% confidence interval of z_Q is determined by

$$Q(z_Q) - Q_{\min} \leq 3.84. \tag{16}$$

Two QTL mapping

If the one-QTL model is not appropriate, then there may be two QTLs, denoted by QTL_1 and QTL_2. Assuming that the positions and effects of QTL_1 and QTL_2 are z_{Q1}, b_1 and z_{Q2}, b_2, respectively, and the effect of interaction between b_1 and b_2 is $b_{1,2}$; then given that the recombination rate between a marker and QTL_1 and that between the marker and QTL_2 are r_1 and r_2, respectively, it can be found that

$$y = b_1 x_1 + b_2 x_2 + b_{1,2} x_{1,2} \qquad (17)$$

where, similar to (7), $y = M_1 - M_2$, $x_1 = 1 - 2r_1$, $x_2 = 1 - 2r_2$, whereas $x_{1,2} = 1 - r_1 - r_2$.

Neglecting the interaction effect (i.e., $b_{1,2} = 0$) this relationship follows a four-parameter nonlinear regression model of the type

$$y_i = b_1 x_{1i}(z_{Q1}) + b_2 x_{2i}(z_{Q2}) + \varepsilon_i \qquad (18)$$

where both $x_{1i}(z_{Q1})$ and $x_{2i}(z_{Q2})$ are determined by (9). Similar to the one-QTL model, parameters in model (18) can be estimated with WLS, and the fitness of the model can be tested with a χ^2-test [here $Q_{\min} \sim \chi^2(n-4)$]. There is no need to test the significance of $\hat{b}_1$ and $\hat{b}_2$ here, because the fact that the one-QTL model is not fit already indicates the existance of more than one QTL. As to the confidence intervals of z_{Q1} and z_{Q2}, both of them can be determined by (16).

In order to reach the point of $Q_{\min}$ quickly, it is necessary to have suitable initial values of z_{Q1} and z_{Q2}. They can be found by means of a t-test or an ANOVA on every marker locus. Theoretically, maximal t or F values would appear on the markers closest to either QTL_1 or QTL_2.

The interaction between the two QTLs should be taken into account when model (18) is not appropriate. If the complete two-QTL model is still unfit, more than two QTLs may exist on the chromosome.

Multi-QTL mapping

If interactions among three or more QTLs are negligible, the two-QTL model can be extended to the case of multiple QTLs, of which the general formula, like (17), is

$$y = \sum_{i=1}^{l} b_i x_i + \sum_{i<j}^{l} b_{ij} x_{ij} \qquad (19)$$

where l is the number of QTLs, and the definitions of other symbols (y, b_i, x_i, x_{ij}) are similar to those in (17). Obviously, there is no difference in principle between multi-QTL mapping and two-QTL mapping. We will not, therefore, discuss them in detail in the present paper.

Examples

One-QTL model

To illustrate the feasibility of joint mapping in the case of only one QTL on a chromosome, let us consider a theoretical example given by Lander and Botstein (1989). In this example, it was assumed that there were 12 chromosomes being tested in an organism, each chromosome was 100 cM long and had six marker loci in total, one every 20 cM from the left to the right. There was a QTL on each of chromosomes 1–5 (Table 1), but no QTL on chromosomes 6–12. The environmental noise had a standard deviation of 1. The sample size was 250. All the simulated data were produced on a computer.

The results are listed in Table 2. As expected, the one-QTL model gives a fit to all the chromosomes. Similar to the results of Lander and Botstein (1989), obtained by interval mapping, every chromosome with a QTL except chromosome 5 has a significant estimate $\hat{b}$, and the significance increases as the QTL's effect increases, as does the reliability of the estimation of the QTL's position. With regard to chromosome 5, although its $\hat{b}$ does not reach the 5% overall sig-nificance level, it is still much greater than those of the chromosomes without QTLs, and will be significant if a 5% nominal significance level is used. Hence, in practice, it is necessary to re-examine such kinds of uncertain chromosomes by further experiments.

Table 1 Positions and effects of QTLs on the first five chromosomes based on a one-QTL model

Chromosome	z_Q (cM)	b	σ	b_0 (b/σ)
1	70	1.50	1.358	1.105
2	49	1.25	1.420	0.880
3	27	1.00	1.468	0.681
4	8	0.75	1.505	0.498
5	30	0.50	1.531	0.327

Table 2 Estimation results for the 12 chromosomes based on a one-QTL model

Chromosome	$Q_{\min}$[a]	$\hat{b}$	$Q(b=0) - Q_{\min}$[b]	$\hat{z}_Q$	95% Confidence interval of z_Q	$\hat{\sigma}$[c]
1	2.499	1.577	154.54*	69	61–77	1.355
2	0.856	1.314	105.73*	51	42–64	1.424
3	2.922	0.988	56.02*	24	2–37	1.488
4	2.555	0.894	38.69*	9	0–38	1.503
5	0.554	0.318	5.64?	22	0–100	1.560
6	2.444	−0.116	0.62	–	–	1.568
7	3.059	−0.198	1.80	–	–	1.568
8	1.336	0.071	0.23	–	–	1.568
9	0.585	−0.219	2.31	–	–	1.568
10	1.496	0.087	0.43	–	–	1.568
11	1.593	0.161	1.45	–	–	1.568
12	4.455	0.213	2.06	–	–	1.568

[a] $\chi^2_{0.05} = 9.488$ ($df = 6 - 2 = 4$)

[b] When $\alpha_0 = 0.05$, $\chi^2_\alpha = \chi^2_{0.00427} = 8.189$. *, significancant; ?, uncertain

[c] Calculated by the formula $\sigma^2 = V - b^2/4$, where V is the total variance

Two-QTL model

In the case of two QTLs on a chromosome, we also use a theroretical example, given by Lander and Botstein (1989). In this example, it was assumed that there were two QTLs located respectively at positions 50 cM and 130 cM from the left end of a chromosome, which was 200 cM long and had 11 marker loci in total, one every 20 cM from the left to the right. Both the QTLs had an effect of 0.9. The sample size was again 250. In addition, we also consider a repulsion-phase situation in which in the above example the effect of the QTL located at 130 cM was changed to -0.9. Simulated data were again produced on a computer.

The results show that the one-QTL model is unfit in these two examples ($Q_{\min} = 60.49$ and 69.54 respectively, $P < 0.001$) and the two-QTL model may be needed. After a t-test on every marker locus (Table 3), it is evident that, on both the coupling-phase chromosome (CPC) and the repulsion-phase chromosome (RPC), there should be two QTLs, one located between 0 and 80 cM, and the other between 100 and 200 cM, but it is not easy to decide suitable values of z_{Q1} and z_{Q2} on the RPC. However, they can be determined by applying the one-QTL model to each of the only-one-QTL-containing regions because there are enough markers within each of them (Table 4).

Table 3 T-test results on the coupling-phase (CPC) and the repulsion-phase (RPC) chromosome based on a two-QTL model

Marker	Position (cM)	*t*-value[a]	
		CPC	RPC
1	0	4.559**	0.693
2	20	3.543**	1.085
3	40	8.629**	2.301*
4	60	8.261**	1.069
5	80	5.474**	2.616**
6	100	5.839**	−0.588
7	120	7.513**	−0.530
8	140	4.806**	−1.853
9	160	6.537**	−1.435
10	180	2.207*	−1.871
11	200	1.283	−0.565

[a] * or **, significant at 5% or 1% level

It is necessary to note that the estimate of a QTL's effect obtained in this way does not reflect the real effect (b) of the QTL but, rather, a mixture (denoted by b') containing a contribution from the other QTL on the same chromosome. Assuming the recombination rate between QTL_1 and QTL_2 is r, it is easy to show that $b'_1 = b_1 + (1-2r)b_2$, $b'_2 = b_2 + (1-2r)b_1$, or

$$b_1 = \frac{b'_1 - (1-2r)b'_2}{4r(1-r)}, \tag{20}$$

$$b_2 = \frac{b'_2 - (1-2r)b'_1}{4r(1-r)}. \tag{21}$$

with (20) and (21), estimates of $\hat{b}_1$ and $\hat{b}_2$ in the examples can be obtained (Table 4). They are indeed closer to their real values.

With the values of z_{Q1} and z_{Q2} provided by the one-QTL model, the results off estimation using the two-QTL model are listed in Table 5. The results shows that the two-QTL model gives a fit to both the CPC and the RPC, and the estimates are also ideal.

It is worth noting that, according to the results in Table 3, in the case of the repulsion phase, the offset in effect of the two QTLs may decrease the t value on each marker locus considerably so as to make the method of t-testing less powerful. In fact, in accordance with the results of t-tests alone, it is impossible to draw the conclusion that there is a QTL with a negative effect at position 130 cM in this example. This reveals the draw-

Table 4 Estimation results in one-QTL-containing regions using the one-QTL model

Chrom.	Region (cM)	$Q_{\min}$	$\hat{z}_Q$	95% confidence interval of z_Q	$\hat{b}'$	$Q(b'=0) - Q_{\min}$	$\hat{b}$
CPC	0–80	3.575[a]	51	44–57	1.414	203.21	1.196
	100–200	8.041[b]	125	111–132	1.248	154.46	0.976
RPC	0–80	1.582[a]	44	27–59	0.616	43.45	0.761
	100–200	1.802[b]	134	122–154	−0.753	66.59	−0.879

[a] $\chi^2_{0.05} = 7.815\ (df = 5-2 = 3)$

[b] $\chi^2_{0.05} = 9.488\ (df = 6-2 = 4)$

Table 5 Estimation results using the two-QTL model

Chrom.	$Q_{\min}$	$\hat{z}_{Q1}$	$\hat{z}_{Q2}$	95% confidence interval		$\hat{b}_1$	$\hat{b}_2$
				z_{Q1}	z_{Q2}		
CPC	11.803	47	129	37–55	118–154	1.129	0.919
RPC	10.928	49	130	35–70	121–147	0.882	−0.980

[a] $\chi^2_{0.05} = 14.067 (df = 11-4 = 7)$

backs of the approaches studying single markers one-at-a-time.

Discussion

In the introduction we have mentioned some problems concerning the method of interval mapping. It is obvious, however, that in the joint mapping method these problems disappear. Firstly, since only moments are used in joint mapping, it is unnecessary to know the exact theoretical form of the quatitative trait distribution in the experimental population. Therefore, the approach is widely applicable. Secondly, the threshold for significance testing in joint mapping can be determined conveniently and the statistical confidence intervals of QTL positions can also be easily obtained. Thirdly, no complex mathematical knowledge is needed to understand the principle of joint mapping and the procedure of computation is also simple. Therefore, it is easy to write a computer program and very little computer time is required. It is especially useful when multiple QTLs are involved and/or there are interactions between them on a chromosome. Such complicated cases may make the use of interval mapping impractical (Haley and Knott 1992). Additionally, joint mapping is also very powerful and efficient. In theory,, its power is ultimited if there are as many marker loci as needed. We will discuss this matter in another paper.

A factor possibly limiting the reliability of joint mapping might come from the approximation of Haldane's map function, of which the hypothesis that there is no interference among single exchanges may not always be true. However, the numbers of double and multiple exchanges are generally much smaller than those of the related single exchanges and their standard errors. It seems, therefore, that even though there may be complete inteference, errors caused by the approximation of the map function can still be neglected compared with those caused by sampling.

In addition, although only the first generation backcross population was discussed in this paper, in view of the similarities in population genetic structure, the results obtained here can be applied to populations of doubled haploid lines and single chromosome homozygous recombinant lines directly, provided the heterozygous genotype (A_1A_2) in BC_1 is replaced by the corresponding homozygous genotype (A_2A_2). For the same reason, the results are also basically suitable for a population of recombinant inbred lines as long as the parameter of recombination rate (r) in the models is substituted by R $[= 2r/(1 + 2r)]$ (Simpson 1989). As for the F_2 population, the basic principle of joint mapping is, as we will discuss elsewhere, again applicable.

References

Beckmann JS, Soller M (1986) Restriction fragment length polymorphisms in plant genetic improvement. Oxford Surveys Plant Mol Cell Biol 35:97–246

Carbonell EA, Gering TM, Balansard E, Asins MJ (1992) Interval mapping in the analysis of nonadditative quantitative trait loci. Biometrics 48:305–315

Ellis THN (1986) Restriction fragment length polymorphism markers in relation to quantitative characters. Theor Appl Genet 72:1–2

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

Hamilton D (1986) Confidence regions for parameter subsets in nonlinear regression. Biometrika 73:57–64

Johansen S (1984) Functional relations, random coefficients, and nonliear regression with application to kinetic data. Springer-Verlag, New York, pp 77–80

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

Li WM, Wu WR, Lu HR (1993) A method of detecting linkage between quantitative trait loci and genetic marker and its application in wheat. Acta Agron Sinica (in press)

Luo ZW, Kearsey MJ (1989) Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. Heredity 63:401–408

Luo ZW, Kearsey MJ (1992) Interval mapping of quantitative trait loci in an F_2 population. Heredity 69:236–242

Mather K, Jinks JL (1971)

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

Snape JW, Law CN, Parker BB, Worland AJ (1985) Genetic analysis of chromosome 5A of wheat and its influence on important agronomic characters. Theor Appl Genet 71:518–528

Tanksley SD, Young ND, Paterson AH, Bonierbale MW (1989) RFLP mapping in plant breeding: new tools for an old science. Biotechnology 7:257–264

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

Wu WR, Li WM, Lu HR (1991) A moment approach for estimating linkage values between QTLs and RFLP loci and other parameters using recombinant inbred lines (abstract, in Chinese). Genetic Research in China. Symp 4th Congr Genetics Society of China pp 215–216