Theor Appl Genet (1996) 93: 1156–1160

Wei-Ren Wu · Wei-Ming Li

Joint mapping of quantitative trait loci using F_2 populations

Received: 15 January 1996 / Accepted: 17 May 1996

Abstract In this paper, the theory of joint mapping of quantitative trait loci is extended to F_2 populations. Two independent regression equations, related to the additive and dominance effects respectively, are derived. Therefore, there are three alternative strategies for mapping QTLs, called additive-based mapping (ABM), dominance-based mapping (DBM) and additive-dominance-based mapping (ADBM). Simulation results have shown that ADBM is the most appropriate in most situations.

Key words QTL · Joint mapping · F_2 population · Genetic marker · Linkage

Introduction

A method for mapping quantitative trait loci (QTLs) in a backcross population, called joint mapping (JM), was proposed in previous papers (Wu and Li 1994, 1996). The method can be applied directly or with slight modification to doubled-haploid lines (DH), recombinant inbred lines (RIL), or other populations with a similar genetic structure. The basic principles of the method are also applicable to F_2 populations. In fact, Kearsey and Hyne (1994) independently proposed a method similar to ours suitable for F_2 populations. But they only discussed the case involving one QTL on a chromosome. In addition, there are still problems in model fitting and model testing in their method due to the correlated associations of linked markers. These problems can be solved by using the general least square method instead of the conventional least square method to fit the regression model (Wu and Li 1996). In the present paper, we extend our theory to F_2 populations.

Communicated by J. W. Snape

W.-R. Wu (✉) · W.-M. Li
Department of Agronomy, Fujian Agricultural University, Fuzhou, Fujian, PRC

Theory

Suppose that an F_2 is derived from two pure-line parents, P_1 and P_2. For any polymorphic locus, alleles from P_1 and P_2 will be indicated by subscripts 1 and 2, respectively. Consider a chromosome with n markers. Let M_{11i}, M_{12i} and M_{22i} denote the means of the three genotypes, A_1A_1, A_1A_2 and A_2A_2, of the ith marker, respectively. If there is a QTL (with alleles Q_1 and Q_2) on the chromosome, we can obtain the following two orthogonal or independent equations (Kearsey and Hyne 1994):

$$M_{12i} - (M_{11i} + M_{22i})/2 = (1 - 2r_i)^2\, d \tag{1}$$

$$(M_{11i} + M_{22i})/2 = (1 - 2r_i)\, a \tag{2}$$

where a and d are the additive and the dominance effects of the QTL (i.e. $Q_1Q_1 = a$, $Q_1Q_2 = d$ and $Q_2Q_2 = -a$), respectively; and r_i is the recombination frequency between the ith marker and the QTL.

Equations (1) and (2) can be extended to the case of multiple (say m) linked QTLs. Following the system of Mather and Jinks (1982), and assuming no exchange interference, it can be shown that

$$(M_{11i} - M_{22i})/2 = \sum_{k=1}^{m}(1-2r_{ik})\, a_k + 2\sum_{k\neq 1}^{m}(1-2r_{ik})\, r_{il}\,(1-r_{il})\, ad_{kl} \tag{3}$$

$$M_{12i} - (M_{11i} - M_{22i})/2 = \sum_{k=1}^{m}(1-2r_{ik})^2\, d_k - \sum_{k\neq 1}^{m}(1-2r_{ik})\,(1-2r_{il})\, aa_{kl} + \sum_{k\neq 1}^{m}[1-2r_{ik}(1-2r_{ik}) - 2r_{il}(1-2_{il})]\, dd_{kl} \tag{4}$$

where subscripts k and l indicate the kth and the lth QTLs, respectively; and aa, ad and dd denote the epistatic effects of $a \times a$, $a \times d$ and $d \times d$, respectively.

For the sake of simplicity, in the following discussion epistatic effects among linked QTLs will be neglected. In this case, $(1/2)(M_{11i} - M_{22i})$ and $M_{12i} - (1/2)(M_{11i} + M_{22i})$

are related only to additive effects and dominance effects of linked QTLs, respectively, and both (3) and (4) can be expressed as the following regression model:

$$y_i = \sum_{k=1}^{m} b_k x_{ik} + e_i \qquad (i = 1,\ldots,n) \tag{5}$$

where, for (3), $y_i = (1/2)(M_{11i} - M_{22i})$, $x_{ik} = 1 - 2r_{ik}$; and for (4), $y_i = M_{12i} - (1/2)(M_{11i} + M_{22i})$, $x_{ik} = (1 - 2r_{ik})^2$; and e_i is random error, which, according to the central limit theorem, asymptotically follows a normal distribution $N(0, \sigma_i^2)$. Following Haldane's map function (Haldane and Waddington 1931), we find that for (3), $x_{ik} = \exp\{-2|p_i - q_k|\}$ and for (4), $x_{ik} = \exp\{-4|p_i - q_k|\}$, where p_i and q_k are the positions of the ith marker and the kth QTL, respectively (in Morgan units).

Obviously, model (5) (the F_2 model) has the same mathematical form and properties as those of the model for BC populations (the BC model) given in the previous paper (Wu and Li 1996). Therefore, the method of general least squares (GLS) can be used to estimate the parameters. In other words, the residual sum of squares (RSS) is

$$RSS = (Y - XB)' \Sigma^{-1} (Y - XB) \tag{6}$$

where Y is an $(n \times 1)$ vector of y_i s; X is an $(n \times m)$ matrix of x_{ik}s; and B is an $(m \times 1)$ vector of b_ks; $\Sigma = \text{var}(Y)$, the variance matrix of Y, which, for equation (3), can be estimated by (A1) or, more conveniently, by (A9), and for equation (4), can be estimated by (A2) or, more conveniently, by (A11) (see Appendix). Additionally, as we have shown by simulation in the previous paper (Wu and Li 1995), for an m-QTL model, the RSS [denoted as $RSS(m)$] approximately follows a chi-square distribution with n–$2m$ degrees of freedom, i.e., $RSS(m)$ is distributed approximately as $\chi^2(n-2m)$ and, therefore, $RSS(m) - RSS(m+1)$ is distributed as $\chi^2(2)$. These properties can be utilized to find the best model (Wu and Li 1996), that is the optimal (say, m-QTL) model, which should meet the inequalities $RSS(m) < \chi^2_\alpha(n-2m)$, $RSS(m-1) - RSS(m) > \chi^2_\alpha(2)$, and $RSS(m) - RSS(m+1) < \chi^2_\alpha(2)$, where α is the significance level required in accordance with the number of chromosomes being tested (Wu and Li 1996).

In addition, we have found that the coefficient of correlation between y_i and y_j (denoted as ρ_{ij}) is expected to be approximately $\rho_{ij} = 1 - 2r_{ij}$ for (3) and $\rho_{ij} = (1 - 2r_{ij})^2$ for (4) (see Appendix), both of which, under the assumption of Haldane's map function, have the same property as that in BC model as follows:

$$\rho_{ij} = \rho_{i,i+1}\,\rho_{i+1,i+2}\cdots\rho_{j-1,j} \qquad (i < j). \tag{7}$$

This property simplifies the computation of GLS (Wu and Li 1996).

It is noted that the F_2 model consists of two equations, (3) and (4). Hence, there may be three alternative strategies for mapping QTLs. The first strategy is to use (3) to map QTLs but use (4), which is reduced into a linear equation as the positions of QTLs are given by (3), to estimate dominance effects (Kearsey and Hyne 1994). This strategy may be called additive-based mapping (ABM). The second strategy is the opposite of the first one and may be called dominance-based mapping (DBM). The third strategy is to use both (3) and (4) jointly to map QTLs and may be called additive-dominance-based mapping (ADBM). Let *RSSA* and *RSSD* stand for the residual sum of squares of (3) and (4), respectively. Since (3) and (4) are independent, we may define a joint residual sum of squares $RSSJ = RSSA + RSSD$. By minimizing *RSSJ*, we will get GLS estimates of q_k, a_k and d_k ($k = 1,\ldots, m$). As there are $3m$ parameters being estimated using $2n$ nodes of observation (each equation has n nodes), it is expected that the *RSSJ* would approximately follow a chi-square distribution with $2n - 3m$ degrees of freedom; that is, $RSSJ(m)$ would be distributed approximately as $\chi^2(2n - 3m)$, and $RSSJ(m) - RSSJ(m+1)$ would be distributed approximately as $\chi^2(3)$. So the criteria for significance tests of the m-QTL model would be, $RSSJ(m) < \chi^2_\alpha(2n - 3m)$, $RSSJ(m-1) - RSSJ(m) > \chi^2_\alpha(3)$, and $RSSJ(m) - RSSJ(m+1) < \chi^2_\alpha(3)$.

To verify the assumption that $RSSJ(m)$ be distributed approximately as $\chi^2(2n - 3m)$, a simulation study was conducted. In the simulation, a 20-cM long chromosome with five evenly located markers, but without QTLs (i.e., under the null hypothesis), was assumed (for the sake of simplicity in computation, only a very short chromosome was considered, but the simulation results should still be able to reveal the statistical behavior of *RSSJ*). The random error was assumed to follow a standard normal distribution. The sample size was set at 300 individuals and 500 simulations were conducted. The results are given in Table 1, which clearly show that the above assumption is correct.

There is a simple way to test if (3) and (4) are informative for QTL mapping. Note that the total sum of squares of (3) (denoted as *TSSA*) or (4) (denoted as *TSSD*) would approximately follow a central chi-square distribution with n degrees of freedom, if there are no QTLs on the chromosome, or if QTLs exist but their additive (for *TSSA*) or dominance (for *TSSD*) effects (but not both) are zero. Hence, statistically, (3) or (4) would be informative only when $TSSA > \chi^2_\alpha(n)$ or $TSSD < \chi^2_\alpha(n)$. And the larger *TSSA* or *TSSD* is, the more informative (3) or (4) will be. Thus, we are able to find criteria to identify the optimal strategy for mapping QTLs. In principle, if $TSSA > \chi^2_\alpha(n)$ but $TSSD < \chi^2_\alpha(n)$, then ABM may be the optimal. In contrast, if $TSSD > \chi^2_\alpha(n)$ but $TSSA < \chi^2_\alpha(n)$, then DBM may be the optimal, but such a case might be rare in general. If both *TSSA* and *TSSD* are significant, then ADBM may be the best.

Example

It was assumed that a chromosome is 100 cM in length with 11 evenly-spaced markers and three QTLs located at 7, 45 and 82 (cM). Four different effects for the linked QTLs were considered (Table 2), i.e., case 1: no dominance; case 2: complete dominance; case 3: over-dominance; and case 4: mixture. The residual error (caused by environmental variation and the segregation of QTLs located on other

Table 1 Results of simulated sampling of RSSJ (see text). MRSSJ: mean of RSSJ; VRSSJ: variance of RSSJ; exp: expected, according to the theorectical (chi-square) distribution; sam: sampled by simulation; *df*: degree of freedom; *P*: upper probability of χ^2

Model		MRSSJ	VRSSJ	Goodness of fit		
				χ^2	*df*	*P*
Null-QTL	exp	10	20			
	sam	10.191	21.821	4.768	11	0.942
One-QTL	exp	7	14			
	sam	6.923	13.883	4.440	9	0.880

Table 2 Real positions (cM) and effects of QTLs in the simulated example

Case	Parameter	QTL1	QTL2	QTL3
1–4	*q*	7	45	82
1	*a*	0.8	–0.8	0.8
	d	0.0	0.0	0.0
	h^2 (%)[a]	16.33	16.33	16.33
2	*a*	0.8	–0.8	0.8
	d	0.8	0.8	–0.8
	h^2 (%)	19.67	19.67	19.67
3	*a*	0.4	–0.4	0.4
	d	0.8	0.8	–0.8
	h^2 (%)	13.95	13.95	13.95
4	*a*	0.8	–0.8	0.4
	d	0.0	0.8	–0.8
	h^2 (%)	15.69	23.53	11.76

[a] h^2: broad heritability

Table 3 TSSA and TSSD of each case in the simulated example. TSSA, TSSD~$\chi^2(11)$; *P*: upper probability

Case	TSSA	*P*	TSSD	*P*
1	66.071	<0.0001	14.397	0.2118
2	64.198	<0.0001	61.349	<0.0001
3	21.785	0.0261	125.128	<0.0001
4	49.823	<0.0001	49.557	<0.0001

chromosomes) was assumed to follow a normal distribution with the variance set equal to 1. A sample of genotypes of all the loci with 300 individuals was generated. For the sake of comparison, a sample of quantitative trait phenotypes for each case was generated based on the same sample of genotypes.

The *TSSA* and *TSSD* of each case were first calculated (Table 3). We see that *TSSA* and *TSSD* reflect the relative importance of additive effects and dominance effects of the QTLs, respectively. So, in accordance with the criteria mentioned above, the possible optimal strategy for each case can be identified, i.e., ABM for case 1 and ADBM for cases 2–4. But in order to have an overall comparison among the strategies, all of them were used to analyze the data, except that DBM was not applied to case 1 because *TSSD* was not significant there (Table 4).

The results show that the optimal strategies are as expected. It is noted that ADBM seems always to be suitable, even to the case (case 1) where dominance effects are zero. This is obvious because *RSSJ* is the sum of *RSSA* and *RSSD*. *RSSD* (or *RSSA*) will make only a small contribution to *RSSJ* when dominance (or additive) effects of linked QTLs are zero. In this case, therefore, there will be no great difference between ADBM and ABM (or DBM). For a mixed case (case 4), which may be more frequently met in practice, ADBM shows an apparent advantage over ABM and DBM. Hence, generally, ADBM will be the best method.

Discussion

With the two previous papers (Wu and Li 1994, 1996) and the present one, we have now basically constructed the theoretical framework of the JM method. Obviously, the most desirable features of JM are that it integrates separate tests of markers on a chromosome into one analysis and has a flexible model suitable for mapping different number of linked QTLs. Therefore, it may, at least in some cases, achieve comparatively high statistical powers (Wu and Li 1996) and seldom suffer the problem of mapping 'ghost' QTLs, a phenomenon that often occurs in the method of interval mapping (Lander and Botstein 1989) when multiple linked QTLs exist (Martínez and Curnow 1992). In principle, any number of QTLs on a chromosome can be distinguished and precisely mapped by JM as long as there are sufficient markers and a large enough sample size, except for the case when two closely linked QTLs are located in neighbouring marker intervals and have effects in same direction. In such a case, additional markers located between the two QTLs are needed to provide more information.

Appendix: covariance and coefficient of correlation between two observations of the dependent variable in model (5)

Assume that the *i*th and the *j*th markers are linked with a recombination frequency r_{ij}. Let *N* denote the number of individuals of the sample; *n* denote the number of individuals of a genotype; *V* denote the variance of a genotype; and subscripts 11, 12 and 22 indicate genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. Then, according to the result of Appendix 2 of the previous paper (Wu and Li 1996), we find that for equation (3), the covariance between y_i and y_j is

$$\begin{aligned} Cov(y_i,y_j) &= Cov[(M_{11i} - M_{22i})/2, (M_{11j} - M_{22j})/2] \quad \text{(A1)} \\ &= Cov(M_{11i} - M_{11j})/4 - Cov(M_{11i} - M_{22j})/4 \\ &\quad - Cov(M_{22i} - M_{11j})/4 + Cov(M_{22i}, M_{22j})/4 \end{aligned}$$

Table 4 Results of QTL mapping with different strategies. RSS: residual sum of squares; *df*: degree of freedom; *P*: upper probability; *: Significant in chi-square test. That means the reduced model cannot fit the data well and, therefore, the strategy is not suitable for the case

Case	strategy	Parameter	QTL1	QTL2	QTL3	RSS	*df*	*P*
1	ADEM	*q*	13.0	44.0	80.0			
		a	0.728	–0.814	0.616	9.649	13	0.7224
		d	0.312	0.064	–0.202			
	ABM	*q*	12.0	44.0	80.0			
		a	0.709	–0.794	0.615	2.419	5	0.7886
		d	0.297	0.074	–0.202	7.330	8	0.5015
2	ADBM	*q*	10.0	45.0	79.0			
		a	0.669	–0.735	0.770	17.764	13	0.1667
		d	0.606	0.758	–0.650			
	ABM	*q*	11.0	47.0	80.0			
		a	0.676	–0.732	0.767	6.174	5	0.2897
		d	0.657	0.682	–0.614	12.792	8	0.1192
	DBM	*q*	4.0	35.0	78.0			
		a	0.580	–0.549	0.659	22.766	8	0.0037
		d	0.638	0.811	–0.610	6.128	5	0.2940
3	ADBM	*q*	8.0	55.0	76.0			
		a	0.367	–0.303	0.325	14.157	13	0.3629
		d	1.189	0.945	–0.921			
	ABM	*q*	10.0	–	–			
		a	0.333	–	–	10.482	9	0.3129
		d	1.165	–	–	41.712*	10	<0.0001
	DBM	*q*	7.0	54.0	76.0			
		a	0.358	–0.276	0.308	7.677	8	0.4656
		d	1.218	0.927	–0.891	7.279	5	0.2007
4	ADBM	*q*	8.0	43.0	80.0			
		a	0.659	–0.700	0.467	16.168	13	0.2402
		d	0.0313	0.740	–0.744			
	ABM	*q*	9.0	35.0	80.0			
		a	0.772	–0.790	0.409	6.852	5	0.2319
		d	–0.037	0.691	–0.687	12.664	8	0.1240
	DBM	*q*	–	43.0	81.0			
		a	–	–0.356	0.440	34.879*	9	0.0001
		d	–	0.749	–0.774	7.995	7	0.3366

$$= \frac{n_{11i11j}}{4\sqrt{n_{11i}n_{11j}}}\sqrt{\frac{V_{11i}}{n_{11i}}}\sqrt{\frac{V_{11j}}{n_{11j}}} - \frac{n_{11i22j}}{4n_{11i}n_{22j}}\sqrt{\frac{V_{11i}}{n_{11i}}}\sqrt{\frac{V_{22j}}{n_{22j}}}$$

$$- \frac{n_{22i11j}}{4\sqrt{n_{22i}n_{11j}}}\sqrt{\frac{V_{22i}}{n_{22i}}}\sqrt{\frac{V_{11j}}{n_{11j}}} + \frac{n_{22i22j}}{4\sqrt{n_{22i}n_{22j}}}\sqrt{\frac{V_{22i}}{n_{22i}}}\sqrt{\frac{V_{22j}}{n_{22j}}}$$

and for equation (4),

$$\begin{aligned} Cov(y_i, y_j) &= Cov[M_{12i} - (M_{11i}+M_{22i})/2, M_{12j} \\ &\quad - (M_{11j} + M_{22j})/2] \\ &= Cov(M_{11i}, M_{11j})/4 - Cov(M_{11i}, M_{12j})/2 \\ &\quad + Cov(M_{11i}, M_{22j})/4 - Cov(M_{12i}, M_{11j})/2 \\ &\quad + Cov(M_{12i}, M_{12j}) - Cov(M_{12i}, M_{22j})/2 \\ &\quad + Cov(M_{22i}, M_{11j})/4 - Cov(M_{22i}, M_{12j})/2 \\ &\quad + Cov(M_{22i}, M_{22j})/4 \end{aligned} \tag{A2}$$

$$= \frac{n_{11i11j}}{4\sqrt{n_{11i}n_{11j}}}\sqrt{\frac{V_{11i}}{n_{11i}}}\sqrt{\frac{V_{11j}}{n_{11j}}} - \frac{n_{11i12j}}{2\sqrt{n_{11i}n_{12j}}}\sqrt{\frac{V_{11i}}{n_{11i}}}\sqrt{\frac{V_{12j}}{n_{12j}}}$$

$$+ \frac{n_{11i22j}}{4\sqrt{n_{11i}n_{22j}}}\sqrt{\frac{V_{11i}}{n_{11i}}}\sqrt{\frac{V_{22j}}{n_{22j}}} - \frac{n_{12i11j}}{2\sqrt{n_{12i}n_{11j}}}\sqrt{\frac{V_{12i}}{n_{12i}}}\sqrt{\frac{V_{11j}}{n_{11j}}}$$

$$+ \frac{n_{12i12j}}{\sqrt{n_{12i}n_{12j}}}\sqrt{\frac{V_{12i}}{n_{12i}}}\sqrt{\frac{V_{12j}}{n_{12j}}} - \frac{n_{12i22j}}{2\sqrt{n_{12i}n_{22j}}}\sqrt{\frac{V_{12i}}{n_{12i}}}\sqrt{\frac{V_{22j}}{n_{22j}}}$$

$$+ \frac{n_{22i11j}}{4\sqrt{n_{22i}n_{11j}}}\sqrt{\frac{V_{22i}}{n_{22i}}}\sqrt{\frac{V_{11j}}{n_{11j}}} - \frac{n_{22i12j}}{2\sqrt{n_{22i}n_{12j}}}\sqrt{\frac{V_{22i}}{n_{22i}}}\sqrt{\frac{V_{12j}}{n_{12j}}}$$

$$+ \frac{n_{22i22j}}{4\sqrt{n_{22i}n_{22j}}}\sqrt{\frac{V_{22i}}{n_{22i}}}\sqrt{\frac{V_{22j}}{n_{22j}}}.$$

(A1) and (A2) are general formulae for calculating $Cov(y_i,y_j)$. They can be simplified. In accordance with genetic theory, it is expected that:

$$n_{11i} = n_{12i}/2 = n_{22i} = n_{11j} = n_{12j}/2 = n_{22j} = N/4 \quad \text{(A3)}$$

$$n_{11i11j} = n_{22i22j} = (1 - r_{ij})^2\, N/4 \quad \text{(A4)}$$

$$n_{11i22j} = n_{22i11j} = r_{ij}^2\, N/4 \quad \text{(A5)}$$

$$n_{11i12j} = n_{12i11j} = n_{12i22j} = n_{22i12j} = r_{ij}(1 - r_{ij})\, N/2 \quad \text{(A6)}$$

$$n_{12i12j} = [1 - 2r_{ij}(1 - r_{ij})]\, N. \quad \text{(A7)}$$

In addition, if epistatic effects are neglected, we find that for the ith marker,

$$V_{12i} = \sigma^2_{res} + \sum_{k=1}^{m} 2r_{ik}(1-r_{ik})\, a_k^2 + \sum_{k=1}^{m} 2r_{ik}(1-r_{ik})\,[1-2r_{ik}(1-r_{ik})]\, d_k^2$$

$$V_{11i} = V_{12i} - \Delta V_i$$

$$V_{22i} = V_{12i} + \Delta V_i$$

where σ_{res^2} is the residual variance (consisting of environmental variance and residual genetic variance caused by QTLs located on other chromosomes), r_{ik}, a_k and d_k take the same meanings as in (3) and (4), and

$$\Delta V_i = \sum_{k=1}^{m} 4r_{ik}(1-r_{ik})(1-2r_{ik})\, a_k\, d_k.$$

Generally speaking, in most, if not all, cases, ΔV_i is much smaller than V_{12i} in an F_2 population. Hence, approximately,

$$V_{11i} = V_{12i} = V_{22i} = V_i. \quad \text{(A8)}$$

Thus, substituting (A3)–(A8) into (A1) and (A2), we get concise formulae for $Cov(y_i,y_j)$ and, therefore, of the coefficient of correlation between y_i and y_j (denoted as r_{ij}). Namely, for equation (3),

$$Cov(y_i, y_j) = (1-2r_{ij}) \sqrt{\frac{2V_i}{N}} \sqrt{\frac{2V_j}{N}} \quad \text{(A9)}$$

$$\rho_{ij} = 1 - 2r_{ij} \quad \text{(A10)}$$

and for equation (4)

$$Cov(y_i, y_j) = (1-2r_{ij})^2 \sqrt{\frac{4V_i}{N}} \sqrt{\frac{4V_j}{N}} \quad \text{(A11)}$$

$$\rho_{ij} = (1 - 2r_{ij})^2. \quad \text{(A12)}$$

References

Haldane JBS, Waddington CH (1931) Inbreeding and linkage. Genetics 16:357–374

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

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

Mather K, Jinks JL (1982) Biometrical genetics, 3rd edn. University Press, Cambridge

Wu WR, Li WM (1994) A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. Theor Appl Genet 89:535–539

Wu WR, Li WM (1996) Model fitting and model testing in the method of joint mapping of quantitative trait loci. Theor Appl Genet 92:477–482