Theor Appl Genet (1994) 87:1021–1027

Detection of additive and dominance effects of QTLs in interval mapping of F_2 RFLP data

T. Hayashi, Y. Ukai

Department of Agrobiology, Faculty of Agriculture, The University of Tokyo, Yayoi 1-1-1, Bunkyo, Tokyo 113, Japan

Received: 26 June 1992 / Accepted: 2 February 1993

Abstract. A model for interval mapping of quantitative trait loci (QTLs) in an F_2 population using genetic markers such as restriction fragment length polymorphisms (RFLPs) is proposed. Based on this model the log-likelihood ratio statistic is divided into two useful statistics which can be used to detect additive and dominance effects of QTLs, separately. The properties of the two statistics were investigated for the theoretical construction of critical regions to test the gene action of QTLs. The model was applied to 1000 simulated-sets of 200 F_2 individuals.

Key words: RFLPs – Interval mapping – Genetic effect

Introduction

As the number of genetic markers available has increased with the advent of restriction fragment length polymorphisms (RFLPs), several methods to locate quantitative trait loci (QTLs) and to estimate the effects of QTLs using genetic markers have been developed. These methods are mainly based on the maximum likeli-hood technique including an analysis of variance (Soller and Brody 1976; Tanksley et al. 1982; Weller 1986; Edwards et al. 1987; Lander and Botstein 1989; Paterson et al. 1991; Carbonell et al. 1992; Luo and Kearsey 1992). Of these, interval mapping, as proposed by Lander and Botstein (1989), provides the most accurate mapping of QTLs. In this method a region between flanking markers on a chromosome can be scanned to detect a QTL and the probable position of that QTL is given as an interval of the chromosome giving a peak of LOD score. This method is particularly effective when a number of markers are available, and it has been applied both to backcross data (Paterson et al. 1988) and F_2 data (Paterson et al. 1991). Lander and Botstein (1989) first suggested that the method could be applied to F_2 data as well as backcross data. However, the distributional properties of test statistics for F_2 data, on the basis of which the critical region is constructed, are more complicated than for backcross data. For example, log-likelihood ratio statistics given by modifying LOD scores on a series of RFLP markers can not be asymptotically regarded as a typical χ^2 process with two degrees of freedom ($2\,df$) but rather as an intractable stochastic process, for F_2 data; in contrast those in backcross data are a typical χ^2 process with $1\,df$ asymptotically. Paterson et al. (1991) constructed the critical region for the LOD score in F_2 data assuming purely additive gene action. Interval mapping methods for F_2 data have been discussed by Carbonell et al. (1992) and Luo and Kearsey (1992), but no theoretical basis for constructing critical regions for LOD scores for F_2 data has been provided.

In this paper, we propose a new model for the analysis of QTLs for F_2 data. The parameterization of the genetic effects of QTLs is slightly different from that in the model adopted by Carbonell et al. (1992). We reconstruct the statistical procedure for detecting QTLs and estimating their effects. New statistics to detect QTLs and to estimate their gene action are also proposed.

Model for analysis of QTLs in an F_2 progeny

F_2 progenies are produced by an F_1 derived from a cross of two inbred lines, p_1 and p_2. Each F_2 progeny is

Communicated by G. E. Hart
Correspondence to: T. Hayashi

scored for both a quantitative trait and a number of genetic markers which are codominant, such as RFLPs. Generally the analysis of QTLs includes two steps. First, the presence of a QTL near a genetic marker is tested. Second, if the existence of a QTL is suggested, the position and effect of the QTL are estimated. We express the position of a marker or a QTL on a chromosome as the distance measured by genetic map units from the left and marker. A putative QTL (Q) is located at the position s between two flanking markers A and B. We shall use r_1, r_2 and $r_{1,2}$ to indicate recombination values over the intervals $Q-A$, $Q-B$, and $A-B$, respectively. Alleles of loci Q, A and B from a parent P_1 are denoted by Q_1, A_1 and B_1 and those from P_2 by Q_2, A_2 and B_2, respectively. The genotype of the F_1 is $A_1Q_1B_1/A_2Q_2B_2$. F_2 individuals can be divided into nine classes according to the marker genotypes, i.e., $A_1A_1B_1B_1$, $A_1A_1B_1B_2$, $A_1A_1B_2B_2$, $A_1A_2B_1B_1$, $A_1A_2B_1B_2$, $A_1A_2B_2B_2$, $A_2A_2B_1B_1$, $A_2A_2B_1B_2$ and $A_2A_2B_2B_2$[1]. The $A_1A_2B_1B_2$ class consists of individuals of genotypes A_1B_1/A_2B_2 and A_1B_2/A_2B_1. These nine marker classes are referred to as classes 1–9, respectively.

Suppose that a total of N progeny are sampled and that the number of individuals which fall in class i is $n_i(\sum_{i=1}^{9} n_i = N)$. Let the jth observation of class i be denoted by y_{ij} ($j = 1, 2, \ldots n_i$ for $i = 1, 2, \ldots 9$). When Q is located at the position s and between two markers A and B y_{ij} is described by the following model;

$$y_{ij} = \mu + z'_{ij}g + \varepsilon_{ij} \quad (j = 1, 2, \ldots n_i \text{ for } i = 1, 2, \ldots 9), \tag{1}$$

where μ is the mid-parental value of the contributions of Q added to the mean contributions of the other QTLs and $g = (a, d, -a)'$ is a vector of effects of Q, where a, d and $-a$ denote the genotypic effects of the genotypes Q_1Q_1, Q_1Q_2 and Q_2Q_2, respectively, measured from the mid-parental value. $z_{ij} = (z_{ij1}, z_{ij2}, z_{ij3})'$ is a random vector indicating genotypes of Q of the jth individual of class i, which equals $e_1 = (1, 0, 0)'$, $e_2 = (0, 1, 0)'$ or $e_3 = (0, 0, 1)'$ with probabilities p_{i1}, p_{i2} or p_{i3} corresponding to the genotypes Q_1Q_1, Q_1Q_2 and Q_2Q_2, respectively. p_{i1}, p_{i2} and p_{i3}, which are conditional probabilities given genotypes of the flanking markers ($p_{i1} + p_{i2} + p_{i3} = 1$), can be expressed by a function of the recombination values, r_1, r_2 and $r_{1,2}$. Table 1 shows frequencies of the genotypes of Q of the gametes generated by F_1 individuals given the genotypes of flanking markers A and B. p_{i1}, p_{i2} and p_{i3} are given for each i as in Table 2, where $r_{1,2}$ is the probability that recombination occurs simultaneously in both intervals, $A-Q$ and $Q-B$. If Q is located exactly at A, then $r_1 = 0$ and $z_{1j} = z_{2j} = z_{3j} = e_1, z_{4j} = z_{5j} = z_{6j} = e_2$, and $z_{7j} = z_{8j} = z_{9j} = e_3$. ε_{ij} is a residual and is assumed to be a random normal variable with mean 0 and unknown variance σ^2, which includes environmental variance and residual genotypic variance contributed by the QTLs other than Q.

The parameters are $\theta = (\mu, a, d, \sigma^2)'$ and s, the position of Q. The joint distribution of y_{ij} and z_{ij} can be

Table 1. Conditional frequencies of QTL genotypes generated by F_1 individuals given flanking markers

Genotype of gamete	Frequency of gamete given flankling markers[a]
$A_1Q_1B_1$	$q_1 = (1 - r_1 - r_2 + r_{12})/(1 - r_{1,2})$
$A_1Q_2B_1$	$q_2 = r_{12}/(1 - r_{1,2})$
$A_1Q_1B_2$	$q_3 = (r_2 - r_{12})/r_{1,2}$
$A_1Q_2B_2$	$q_4 = (r_1 - r_{12})/r_{1,2}$
$A_2Q_1B_1$	q_4
$A_2Q_2B_1$	q_3
$A_2Q_1B_2$	q_2
$A_2Q_2B_2$	q_1

[a] r_1: recombination value between A and Q
r_2: recombination value between B and Q
r_{12}: probability of recombination between A and Q and between B and Q on the same gamete
$r_{1,2}$: recombination value between A and B
It is noted that $r_{1,2} = r_1 + r_2 - 2r_{12}$

Table 2. Frequencies of QTL genotypes of F_1 individuals given flanking marker genotypes

Flanking marker genotype	Frequency[a]
$A_1A_1B_1B_1$	$p_{11} = q_1^2$
	$p_{12} = q_1q_2$
	$p_{13} = q_2^2$
$A_1A_1B_1B_2$	$p_{21} = q_1q_3$
	$p_{22} = q_1q_4 + q_2q_3$
	$p_{23} = q_2q_4$
$A_1A_1B_2B_2$	$p_{31} = q_3^2$
	$p_{32} = 2q_3q_4$
	$p_{33} = q_4^2$
$A_1A_2B_1B_1$	$p_{41} = q_1q_4$
	$p_{42} = q_1q_3 + q_2q_4$
	$p_{43} = q_2q_3$
$A_1A_2B_1B_2$	$p_{51} = (q_1q_2 + q_3q_4)/2$
	$p_{52} = (q_1^2 + q_2^2 + q_3^2 + q_4^2)/2$
	$p_{53} = (q_1q_2 + q_3q_4)/2$
$A_1A_2B_2B_2$	$p_{61} = q_2q_3$
	$p_{62} = q_1q_3 + q_2q_4$
	$p_{63} = q_1q_4$
$A_2A_2B_1B_1$	$p_{71} = q_4^2$
	$p_{72} = 2q_3q_4$
	$p_{73} = q_3^2$
$A_2A_2B_1B_2$	$p_{81} = q_2q_4$
	$p_{82} = q_1q_4 + q_2q_3$
	$p_{83} = q_1q_3$
$A_2A_2B_2B_2$	$p_{91} = q_2^2$
	$p_{92} = 2q_1q_2$
	$p_{93} = q_1^2$

[a] For q_1, q_2, q_3, and q_4, see Table 1

written as follows (Carbonell et al. 1992);

$$f(y_{ij}, z_{ij}|\theta, s) = \{p_{i1}\phi((y_{ij}-\mu-a)/\sigma)\}^{z_{ij1}} \times \{p_{i2}\phi((y_{ij}-\mu-d)/\sigma)\}^{z_{ij2}} \times \{p_{i3}\phi((y_{ij}-\mu+a)/\sigma)\}^{z_{ij3}}/\sigma, \quad (2)$$

where $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ is the probability density function of the standard normal distribution. Given N samples, the likelihood function from (2) corresponding to the model (1) is

$$L(\theta, s) = \prod_{i=1}^{9}\prod_{j=1}^{n_i} f(y_{ij}, z_{ij}|\theta, s). \quad (3)$$

A procedure of testing whether a QTL is present or not at the position s is based on $L(\theta, s)$ and maximum likelihood estimates (MLEs) of θ and s are given as values of θ and s which maximize $L(\theta, s)$.

The model (1) is similar to that in Carbonell et al. (1992). Their model is here written as

$$y_{ij} = \mu + a x_{ij} + d(1-x_{ij}^2) + \varepsilon_{ij} \times (j = 1, 2, \ldots n_i \text{ for } i = 1, 2, \ldots 9), \quad (4)$$

where x_{ij} is a coded variable taking values of 1, 0, and -1 corresponding to Q_1Q_1, Q_1Q_2 and Q_2Q_2. If Q is located exactly at one of the two flanking markers A and B, the two models (1) and (4) are equivalent. However, if Q is located somewhere between A and B, the coefficient of d in (1) and (4) may make a difference between the two models. In model (4), the coefficient of d may be replaced by any function $g(x_{ij})$ satisfying $g(1) = g(-1) = 0$ and $g(0) = 0$; for example $g(x) = 1 - x^4$. The reason why $g(x_{ij}) = 1 - x_{ij}^2$ should be preferred in (4) is not clear. On the other hand, in model (1) no ambiguity occurs. It should be noted, however, that likelihood arguments are essentially the same whether based on model (1) or (4).

Test for the presence of QTLs

We first consider the null hypothesis $H_0: \theta = \theta_0 = (\mu_0, 0, 0, \sigma_0^2)'$ and the alternative hypothesis $H_1: \theta = (\mu, a, d, \sigma^2)'$ ($a \neq 0$ and/or $d \neq 0$). H_0 implies that there is no QTL contributing to the trait at or near the position s. Under H_0, the likelihood function is

$$L(\theta_0, s) = \prod_{i=1}^{9}\prod_{j=1}^{n_i} f(y_{ij}, z_{ij}|\theta, s) = \prod_{i=1}^{9}\prod_{j=1}^{n_i} p_{i1}^{z_{ij1}} p_{i2}^{z_{ij2}} p_{i3}^{z_{ij3}} \theta(y_{ij}-\mu_0/\sigma_0)/\sigma_0. \quad (5)$$

The likelihood ratio $\Lambda(s)$ given s to test H_0 against H_1 is

$$\Lambda(s) = \max_{\theta\in\Theta_2\cup\Theta_1} L(\theta, s)/\max_{\theta_0\in\Theta_0} L(\theta, s), \quad (6)$$

where Θ_0 and Θ_1 are parameter spaces corresponding to H_0 and H_1, respectively.

If $\Lambda(s)$ is greater than a critical value corresponding to a given significance level, H_0 is rejected and a QTL at or near the position s is declared to be present. While Lander and Botstein (1989) and Paterson et al. (1991) treated $\text{LOD} = \log_{10}\Lambda(s)$, statistical properties of $\lambda(s) = 2\log\Lambda(s)$ under H_0 should be considered here in order to determine a critical region for the test. It is to be noted that z_{ij} is unknown unless Q is located exactly at A or B. For any positions of Q between A and B, $\lambda(s)$ can be calculated with the EM algorithm by regarding z_{ij} as missing observations (Dempster et al. 1977; Carbonell et al. 1992). The statistical property of $\lambda(s)$ obtained with the EM algorithm, however, is not clear. We restrict our consideration to the case where the position of Q is exactly that of a genetic marker A. Then the indicator variable z_{ij} ($i = 1, 2, \ldots 9$) can be known, and $z_{1j} = z_{2j} = z_{3j} = e_1, z_{4j} = z_{5j} = z_{6j} = e_2$ and $z_{7j} = z_{8j} = z_{9j} = e_3$ hold for A_1A_1, A_1A_2 and A_2A_2, respectively. Consider m genetic markers on a chromosome, let the position of the kth marker be s_k ($k = 1, 2, \ldots m$), and assume that $\lambda(s_k)$ is obtained for each k. $\lambda(s_k)$ follows asymptotically a χ^2 distribution with $2df$ under H_0, where $2df$ correspond to the difference in the number of parameters involved in H_0 and H_1, respectively. From well-known results concerning linear regression, $\lambda(s_k)$ can be expressed as

$$\lambda(s_k) = N\log(\hat{\sigma}_0^2/\hat{\sigma}^2),$$

where $\hat{\sigma}_0^2 = \sum_{i=1}^{9}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{..}})^2/N$ and by putting $u_{ij} = u_{ij1} - z_{ij3}$

$$\hat{\sigma}^2 = \Bigg[\sum_{i=1}^{9}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{..}})^2 - \Big\{\sum_{i=1}^{9}\sum_{j=1}^{n_i}(u_{ij}-\overline{u_{..}})(y_{ij}-\overline{y_{..}})\Big\}^2 \Big/ \sum_{i=1}^{9}\sum_{j=1}^{n_i}(u_{ij}-\overline{u_{..}})^2 - \Big\{\sum_{i=1}^{9}\sum_{j=1}^{n_n}(z_{ij2}-\overline{z_{..2}})(y_{ij}-\overline{y_{..}})\Big\}^2 \Big/ \sum_{i=1}^{9}\sum_{j=1}^{n_i}(z_{ij2}-\overline{z_{..2}})^2\Bigg] \Big/ N. \quad (7)$$

If we put

$$\rho(s_k) = \sqrt{N}\sum_{i=1}^{9}\sum_{j=1}^{n_i}(u_{ij}-\overline{u_{..}})(y_{ij}-\overline{y_{..}}) \Big/ \sqrt{\sum_{i=1}^{9}\sum_{j=1}^{n_i}(u_{ij}-\overline{u_{..}})^2 \sum_{i=1}^{9}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{..}})^2}, \quad (8)$$

and

$$\eta(s_k) = \sqrt{N} \sum_{i=1}^{9} \sum_{j=1}^{n_i} (z_{ij2} - \overline{z_{..2}})(y_{ij} - \overline{y_{..}}) \Big/ \sqrt{\sum_{i=1}^{9} \sum_{j=1}^{n_i} (z_{ij2} - \overline{z_{..2}})^2 \sum_{i=1}^{9} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_{..}})^2}, \quad (9)$$

then we can express $\lambda(s_k)$ as

$$\lambda(s_k) = -N \log[1 - \rho(s_k)^2/N - \eta(s_k)^2/N]. \quad (10)$$

It is noted that the value of $z_{ij} = (z_{ij1}, z_{ij2}, z_{ij3})'$ depends on the marker genotypes at the position s_k. For m positions of genetic markers on a chromosome, a series of correlated tests are performed for H_0. The critical value with significance level α satisfies $\text{Prob}[\max \lambda(s_k) > c] = \alpha$. If $\lambda(s_k)$ exceeds c for some k, H_0 is rejected and the presence of a QTL near the kth marker is declared.

Lander and Botstein (1989) showed that the log-likelihood ratio statistics $\lambda(s_k)$ obtained in their model for the analysis of backcross data is a χ^2 process with $1\,df$ under H_0 and constructed a critical region for the LOD score by applying an extreme stochastic process theory to $\lambda(s_k)$. They also suggested in Appendix 5 that a similar argument can be applied to F_2 data, i.e., $\lambda(s_k)$ for F_2 is χ^2 process with $2\,df$ under H_0, from which the critical region can be constructed Paterson et al. (1988), however, modify the statement and point out that the appropriate critical value depends on the corresponding mathematical theory of large deviations of a generalized χ^2 process (Paterson et al. 1991). The problem of obtaining a critical value for $\lambda(s_k)$ for F_2 data remains unsolved. New statistics in plane of $\lambda(s_k)$ will be discussed here to determine the critical region for testing H_0.

If N is sufficiently large, then it follows from (10) that

$$\lambda(s_k) = \rho(s_k)^2 + \eta(s_k)^2. \quad (11)$$

It is shown that, under H_0, $\rho(s_k)$ and $\eta(s_k)$ are approximately independent normal random variables with mean 0 and variance 1 (Appendix 1), $\text{Cov}[\rho(s_k), \rho(s_1)] = 1 - 2r_{k,l}$ and $\text{Cov}[\eta(s_k), \eta(s_l)] = (1 - 2r_{k,l})^2$ (Appendix 2), where $r_{k,l}$ is the recombination value between the kth marker and the lth marker. Since $\text{Cov}[\rho(s_k), \rho(s_l)] \neq \text{Cov}[\eta(s_k), \eta(s_l)]$, $\lambda(s_k)$ $(k = 1, 2, \ldots m)$ is not a typical χ^2 process but a generalized χ^2 process with $2\,df$.

From the above arguments it can be easily shown that $\rho(s_k)$ and $\eta(s_k)$ $(k = 1, 2, \ldots m)$ are independent stationary Gaussian processes under H_0. The joint distribution of $\rho(s_k)$ $(k = 1, 2, \ldots m)$ is m-variate normal with a mean vector $0 = (0, 0, \ldots 0)'$ and a variance-covariance matrix $\Sigma_1 = (\sigma_{ij})$ $(i, j = 1, 2, \ldots m)$, where diagonal elements, σ_{ij}, are 1 and off-diagonal ones, $\sigma_{ij} (i \neq j)$, are $1 - 2r_{i,j}$. $\eta(s_k)$ $(k = 1, 2, \ldots m)$ is also m-variate normal with a mean vector 0 and a variance-covariance matrix Σ_2, whose diagonal elements are 1 and off-diagonal ones are $(1 - 2r_{i,j})^2$.

Instead of $\lambda(s_k)$, we use $\rho(s_k)$ $(k = 1, 2, \ldots m)$ as a statistic to test H_0. This implies that as an alternative hypothesis we adopt H_1': $a \neq 0$ and $d = 0$ in place of H_1. Paterson et al. (1991) use a similar approach but fail to describe the critical region. The critical values of the test, c, corresponding to a significance level α, are given by the joint distribution of $\rho(s_k)$ $(k = 1, 2, \ldots m)$, where $Pr[\max |\rho(s_k)| > c] = \alpha$. If $|\rho(s_k)|$ takes a value over c for some k, H_0 is rejected and the additive effect of a QTL is detected; thus, the presence of a QTL at or near the kth marker is suggested. Then $\eta(s_k)$ $(k = 1, 2, \ldots m)$ are used as statistics to test H_1'), i.e., whether a dominance effect exists or not. It is easily shown that the joint distribution of $\eta(s_k)$ $(k = 1, 2, \ldots m)$ is the same that under H_0, i.e., $\eta(s_k)$ $(k = 1, 2, \ldots m)$ are m-variate normally distributed with a mean 0 and a variance-covariance matrix Σ_2 even under H_1'. The critical region for $\eta(s_k)$ is obtained as in the case of $\rho(s_k)$. $\lambda(s_k)$ can be initially used to pick up the positions where QTLs are likely to be present. Then, we can detect the additive and dominance effects of the QTLs with $\rho(s_k)$ and $\eta(s_k)$, respectively.

The algorithm for obtaining MLEs of θ and s

If the presence of a QTL at the position of some genetic marker is suggested in testing the hypothesis, then estimators of genetic effects and position of the QTL are obtained by the maximum likelihood method. Although the procedure has been discussed by Carbonell et al. (1992), it is described again for model (1) (see also Luo and Kearsey 1992).

Given the location, s, from (2) and (3) the log-likelihood $\log L(\theta, s)$ is

$$\log L(\theta, s) = \sum_{i=1}^{9} \sum_{j=1}^{n_i} \log f(y_{ij}, z_{ij} | \theta, s)$$

$$= \sum_{i=1}^{9} \sum_{j=1}^{n_i} (z_{ij1} \log p_{i1} + z_{ij2} \log p_{i2} + z_{ij3} \log p_{i3})$$

$$- N/2 \log(2\pi\sigma^2) - \sum_{i=1}^{9} \sum_{j=1}^{n_i} [z_{ij1}(y_{ij} - \mu - a)^2$$

$$+ z_{ij2}(y_{ij} - \mu - d)^2 + z_{ij3}(y_{ij} - \mu - a)^2]/(2\sigma^2). \quad (12)$$

In order for θ to maximize $L(\theta, s)$, given s, $\log L(\theta, s)$ is differentiated with respect to the components μ, a, d

and σ^2 and the following equations are solved.

$$\frac{\partial \log L(\theta, s)}{\partial \mu} = \sum_{i=1}^{9} \sum_{j=1}^{n_i} [z_{ij1}(y_{ij} - \mu - a) + z_{ij2}(y_{ij} - \mu - d) + z_{ij3}(y_{ij} - \mu + a)]/\sigma^2 = 0 \quad (13)$$

$$\frac{\partial \log L(\theta, s)}{\partial a} = \sum_{i=1}^{9} \sum_{j=1}^{n_i} [z_{ij1}(y_{ij} - \mu - a) - z_{ij3}(y_{ij} - \mu + a)]/\sigma^2 = 0 \quad (14)$$

$$\frac{\partial \log L(\theta, s)}{\partial d} = \sum_{i=1}^{9} \sum_{j=1}^{n_i} z_{ij2}(y_{ij} - \mu - d)/\sigma^2 = 0 \quad (15)$$

$$\frac{\partial \log L(\theta, s)}{\partial \sigma^2} = -N/(2\sigma^2) + \sum_{i=1}^{9} \sum_{j=1}^{n_i} [z_{ij1}(y_{ij} - \mu - a)^2 + z_{ij2}(y_{ij} - \mu - d)^2 + z_{ij3}(y_{ij} - \mu + a)^2]/(2\sigma^4) = 0 \quad (16)$$

μ, a, d and σ^2 satisfying (13), (14), (15), and (16) simultaneously, and denoted by $\hat{\mu}, \hat{a}, \hat{d}$ and $\hat{\sigma}^2$, respectively, are

$$\hat{\mu} = \left\{ A \sum_{i=1}^{9} \sum_{j=1}^{n_i} u_{ij} y_{ij} - B \sum_{i=1}^{9} \sum_{j=1}^{n_i} [(1 - z_{ij2}) y_{ij}] \right\} \Big/ (A^2 - B^2),$$

$$\hat{a} = \left\{ \sum_{i=1}^{9} \sum_{j=1}^{n_i} u_{ij} y_{ij} - A\hat{\mu} \right\} \Big/ B,$$

$$\hat{d} = \sum_{i=1}^{9} \sum_{j=1}^{n_i} u_{ij2} y_{ij}/(N - B) - \hat{\mu},$$

and

$$\hat{\sigma}^2 = \sum_{i=1}^{9} \sum_{j=1}^{n_i} [z_{ij1}(y_{ij} - \hat{\mu} - \hat{a})^2 + z_{ij2}(y_{ij} - \hat{\mu} - \hat{d})^2 + z_{ij3}(y_{ij} - \hat{\mu} + \hat{a})^2]/N,$$

where

$$u_{ij} = z_{ij1} - z_{ij3}, A = \sum_{i=1}^{9} \sum_{j=1}^{n_i} u_{ij}$$

and

$$B = \sum_{i=1}^{9} \sum_{j=1}^{n_i} (z_{ij1} + z_{ij3}).$$

Again it should be noted that $z_{ij} = (z_{ij1}, z_{ij2}, z_{ij3})'$ is unknown, unless the position of interest, s, coincides with that of a genetic marker. Then the EM algorithm is used to predict z_{ij} in the above equations for obtaining $\hat{\mu}, \hat{a}, \hat{d}$ and $\hat{\sigma}^2$. The predictor $\tilde{z}_{ij} = (\tilde{z}_{ij1}, \tilde{z}_{ij2}, \tilde{z}_{ij3})$ is obtained as follows (Carbonell et al. 1992; Luo and Kearsey 1992);

$$\tilde{z}_{ijk} = E(z_{ijk} | y_{ij}) = p_{ik} \phi((y_{ij} - \mu - \delta_k)/\hat{\sigma}) \Big/ \sum_{k=1}^{3} p_{ik} \phi((y_{ij} - \mu - \delta_k)/\hat{\sigma}) \quad (k = 1, 2, 3),$$

where $\delta_k = a, d$, and $-a$ for $k = 1, 2$, and 3, respectively. This cycle which consists of predicting z_{ij} given $\hat{\mu}, \hat{a}, \hat{d}$ and $\hat{\sigma}^2$ (E step) and obtaining $\hat{\mu}, \hat{a}, \hat{d}$ and $\hat{\sigma}^2$ given $\tilde{z}_{ij}$ (M step) is iterated until the estimates converge. The limit values in the iteration are denoted by $\hat{\theta}(s) = (\hat{\mu}, \hat{a}, \hat{d}, \hat{\sigma}^2)$. The test statistics $\lambda(s)$, $\rho(s)^2$, and $\eta(s)^2$ are calculated for positions s at a regular distance, say 2 cM, and plotted against the chromosome map. If the presence of a QTL is declared in testing H_0, the position at which the graph of $\lambda(s)$ has its maximum is the MLE of the location of a QTL, $\hat{s}$, and $\hat{\theta}(\hat{s})$ is regarded as the MEL of θ.

Critical region and numerical examples

Let us consider a simple case in which three genetic markers A, B and C are available on a chromosome and the data set is composed of N observations for a quantitative trait. Let allele A_1, B_1 and C_1 be from one parent and A_2, B_2 and C_2 from the other parent. Suppose that the interval $A - B$ and $B - C$ are both 25.5 cM, which corresponds to a recombination value of 0.2 applying the Haldane mapping function. In our simulations we set $\mu = 0$, variance of $\varepsilon_{ij}, \sigma^2 = 1.0$, and the number of observations $N = 200$. The effect of the QTL, $g = (a, d, -a)'$, is given according to the conditions we want to simulate. y_{ij} is determined by the genotype of the QTL, the genetic effect of the QTL, g, and the residual error ε_{ij}, which is generated as a normal random number. First, letting $g = (0, 0, 0)'$ we can contruct a critical region for testing the null hypothesis that there is no QTL at or near the genetic markers. While Lander and Botstein (1989) used extended numerical simulation to determine the critical region of the 5% significance level for LOD scores for the different density of markers, it is possible to construct a critical region of any significance level for test statistics proposed here by multiple integration of a multivariate normal density function when the distances between genetic markers are given. We illustrate this as follows.

The test statistic $\rho(s_k)$ ($k = 1, 2, 3$), where s_k indicates the position of the kth marker and $k = 1, 2$, and 3, correspond to A, B, C, respectively, follows a three-variate normal distribution. The critical region corresponding to a significance level α is constructed by utilizing a three-variate normal density function. The critical value c with a significance level α satisfies $\text{Prob}[\max|\rho(s_k)| > c] = 1 - \text{Prob}(|\rho(s_1)| \leq c, |\rho(s_2)| \leq$

$c, |\rho(s_3)| \leq c$). It is shown that

$$\text{Prob}\,[|\rho(s_1)| \leq c, |\rho(s_2)| \leq c, |\rho(s_3)| \leq c]$$
$$= \int_{-c}^{c}\int_{-c}^{c}\int_{-c}^{c} (2\pi)^{-3/2}|\Sigma_1|^{-1/2} \times \exp(-1/2x'\Sigma_1^{-1}x)dx, \quad (17)$$

where $x = (x_1, x_2, x_3)'$, dx means $dx_1dx_2dx_3$ and the elements σ_{ij} of Σ_1 are covariances between $\rho(s_i)$ and $\rho(s_j)(i,j = 1,2,3)$. It is possible to obtain c numerically from (17). Simulation was carried out to evaluate the critical values obtained from (17) for a sample size of $N = 200$. The number of times for $\max|\rho(s_k)|$ taking the value over c of 1000 simulations are eight for a significance level $\alpha = 0.01$ and 52 for $\alpha = 0.05$, respectively. In this situation the values of c obtained numerically with (17) are 2.92 for $\alpha = 0.01$ and 2.35 for $\alpha = 0.05$, respectively. The critical values of the test statistics proposed here are adequate for given critical regions, even if the sample is of modest size, say $N = 200$.

Finally, the graphs of the statistics, $\lambda(s) = 2\log\Lambda(s)$, $\rho(s)^2$, and $\eta(s)^2$ are illustrated for a sample size of 200 progeny when a QTL is present. A QTL is located at 12.55 cM from A and B. Here two cases were simulated, i.e., additive effect with no dominance for the QTL is assumed for case 1, by letting $g = (0.4, 0.0, -0.4)'$, and complete dominance for case 2, by letting $g = (0.4, 0.4, -0.4)'$. The results are shown in Figs. 1 and 2, where the two horizontal lines indicate critical values with $\alpha = 0.01$ and 0.05 for $\rho(s)^2$, i.e., $(2.92)^2 = 8.35$ and $(2.35)^2 = 5.52$, respectively.' The corresponding LOD scores are 1.81 and 1.19, which are much smaller than the critical values of LOD score suggested by Lander and Botstein (1989), i.e., 2.4 (based on a χ^2 process with 1 df), due to a shorter chromosome length (51 cM vs 100 cM) and the small number of markers (three markers) mapped on the chromosome. It is shown that the presence of a QTL is declared by $\rho(s)^2$ exceeding the

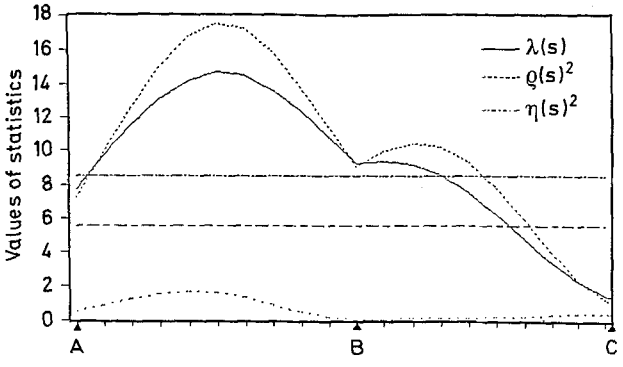

Fig 1. The behavior of statistics $\lambda(s), \rho(s)^2$, and $\eta(s)^2$ along a chromosome map scanned from marker A to marker C for case 1 ($d = 0.4, h = 0.0$). The horizontal lines in the figure indicate critical values for significance levels of 0.01 and 0.05, respectively

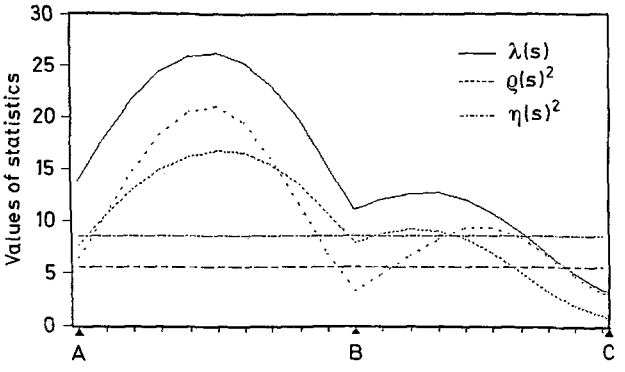

Fig 2. The behavior of statistics $\lambda(s), \rho(s)^2$, and $\eta(s)^2$ along a chromosome map scanned from marker A to marker C for case 2 ($d = h = 0.4$). Two horizontal lines in the figure indicate critical values for significance levels of 0.01 and 0.05, respectively

critical values at A and B for each case and $\eta(s)^2$ sensitive to the dominance effect. As shown in the figures the statistics tend to take higher values between the markers than at the positions of the marker. This tendency is caused by the CM algorithm. Therefore the statistics obtained at the positions of the markers should be used for testing the presence of QTLs. Even if the critical value is exceeded somewhere between the markers, no QTL may be declared when the test statistics show low values at the positions of the flanking markers; for example, no QTL may be suggested in the interval between B and C in Figs. 1 and 2. MLEs of θ and the positions of the QTL are given as those values maximizing $\lambda(s)$ as described above. In simulations a chromosome was scanned from A to C at intervals of 2.55 cM which is 1/10 of the interval between the markers. MLEs obtained in these examples are $\hat{\theta} = (-0.073, 0.359, 0.108, 0.878)$ and $\hat{s} = 12.75$ for case 1 and $\hat{\theta} = (0.073, 0.350, 0.485, 0.892)$ and $\hat{s} = 12.75$ for case 2, while $\theta = (0.0, 0.4, 0.0, 1.0)$, $s = 12.75$ and $\theta = (0.0, 0.4, 0.4, 1.0)$, $s = 12.75$ are given for case 1 and case 2, are respectively.

Discussion and conclusion

The interval mapping method for the analysis of QTLs by Lander and Botstein (1989) has been discussed not only for the backcross data but also for F_2 data (Carbonell et al. 1992; Luo and Kearsey 1992). Though the procedure developed to detect QTLs for backcross data has been applied to F_2 data, no theoretical background exists for constructing the critical region of the LOD score. Unlike backcross data, the distribution of the log-likelihood ratio statistic testing for the presence of a QTL is unknown for F_2 data; thus it is imposible to construct a critical region theoretically. However in a model described here the log-likelihood ratio statistic

can be divided into two useful statistics, $\rho(s)^2$ and $\eta(s)^2$, whose asymptotic statistical properties under the null hypothesis are clear, so that construction of the critical region is possible theoretically, especially for a simple case where only a few markers are under consideration. For the case in which the number of markers available is large, it takes time and cost to calculate a multiple integration such as (17) numerically in order to obtain a critical value, but a critical region may be constructed using simulations. Moreover $\rho(s)$ and $\eta(s)$ are shown to be statistics corresponding to additive and dominance effects of QTLs, respectively. By using these statistics, hypotheses about the gene action of QTLs, such as no dominance and complete dominance, are easily tested. The power of these statistics to detect QTLs under various conditions, for example, sample size and heritability, is under investigation and the results will be discussed elsewhere.

Appendix 1

We show that, for each k, $\rho(s_k)$ and $\eta(s_k)$ are independent normal random variates with mean 0 and variance 1 asymptotically, i.e., when N becomes infinite, H_0. Letting a marker located at s_k be denoted by A, $z_{ij} = (z_{ij1}, z_{ij2}, z_{ij3})$ is e_1, e_2, or e_3 corresponding to the genotypes A_1A_1, A_1A_2 and A_2A_2 with probabilities 1/4, 1/2, and 1/4, respectively. Thus, by means of the law of large numbers, $\overline{u_{..}} = \overline{z_{..1}} - \overline{z_{..3}}$ and $\overline{z_{..2}}$ converge to 0 and 1/2, respectively, with probability 1. For a sufficiently large N, $u_{ij} - \overline{u_{..}}$ takes values of 1, 0, and -1 with frequencies $N/4$, $N/2$, and $N/4$, respectively. Therefore we can write the numerator of the second term of the right-hand side of (8) as

$$\sum_{i=1}^{9}\sum_{j=1}^{n_i}(u_{ij}-\overline{u_{..}})(y_{ij}-\overline{y_{..}}) = \sum{}'(y_{ij}-\overline{y_{..}}) - \sum{}''(y_{ij}-\overline{y_{..}}),$$

where $\sum'$ and $\sum''$ indicate the summation over the set of (i, j) in which $u_{ij} - \overline{u_{..}}$ takes 1 and -1, respectively. Under H_0, y_{ij} follows an independent and identical normal distribution with mean a_0 and variance σ_0^2 for all i and j, so $y_{ij} - \overline{y_{..}}$ are considered to be sampled from a normal distribution with mean 0 and variance σ_0^2 because by means of the law of large numbers $y_{..}$ converges to μ_0 with a probability of 1. Therefore, $\sum_{i=1}^{9}\sum_{j=1}^{n_i}(u_{ij} - \overline{u_{..}})(y_{ij} - \overline{y_{..}})/\sqrt{N}$ is normally distributed with mean 0 and variance $\sigma_0^2/2$. It is also shown by means of the law of large numbers that $\sum_{i=1}^{9}\sum_{j=1}^{n_i}(u_{ij} - \overline{u_{..}})^2$ and $\sum_{i=1}^{9}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{..}})^2$ converge to 1/2 and σ_0^2, so that it is proven that $\rho(s_k)$ defined by (8) is a normal random variable with mean 0 and variance 1. By a similar argument it can be shown that $\eta(s_k)$ is normally distributed with mean 0 and variance 1. Moreover, covariance of $\rho(s_k)$ and $\eta(s_k)$ is 0, since coefficients of $y_{ij} - \overline{y_{..}}$ in (8) and (9), $u_{ij} - \overline{u_{..}}$ and $z_{ij2} - \overline{z_{..2}}$, respectively, are orthogonal. That is, for a sufficiently large N, $u_{ij} - \overline{u_{..}} = 1, 0$, and -1 and $z_{ij2} - \overline{z_{..2}} = -1/2, 1/2$, and $-1/2$ corresponding to the genotype A_1A_1, A_1A_2 and A_2A_2 if marker A is of interest.

Appendix 2

$\mathrm{Cov}[\rho(s_k), \rho(s_l)]$ is treated first. The two markers located at s_k and s_l are denoted by A and B, respectively. For each of N progeny, the indicator variable $z = (z_1, z_2, z_3)$ is given at s_k and s_l corresponding to the marker genotypes (subscripts i and j are omitted for simplicity). Let $u = z_1 - z_3$ at s_k be denoted by u_k and that at s_l be u_l, respectively. The frequencies of the genotype A_1A_1, A_1A_2, and A_2A_2 at A are 1/4, 1/2, and 1/4, respectively, and the conditional frequencies of the genotypes B_1B_1, B_1B_2 and B_2B_2 at B, given the genotypes at A, are $(1 - r_{k,l})^2$, $2r_{k,l}(1 - r_{k,l})$, and $r_{k,l}^2$ for A_1A_1, $r_{k,l}(1 - r_{k,l})$, $(1 - r_{k,l})^2 + r_{k,l}^2$, and $r_{k,l}(1 - r_{k,l})$ for A_1A_2, and $r_{k,l}^2$, $2r_{k,l}(1 - r_{k,l})$, and $(1 - r_{k,l})^2$ for A_2A_2, respectively, where $r_{k,l}$ is a recombination value between A and B. By taking account of $u_k = 1, 0$, and -1 for A_1A_1, A_1A_2, and A_2A_2, respectively, and $u_l = 1, 0$, and -1 for B_1B_1, B_1B_2 and B_2B_2, respectively, it is clear that $\mathrm{Cov}(u_k, u_l) = 1 - 2r_{k,l}$. Therefore from (8) it is easily shown that $\mathrm{Cov}[\rho(s_k), \rho(s_l)] = 1 - 2r_{k,l}$ holds approximately by using arguments similar to those of Appendix 1. By the same consideration of z_2, $\mathrm{Cov}[\eta(s_k), \eta(s_l)] = (1 - 2r_{k,l})^2$ can be proved.

References

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

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc 39:1–38

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

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

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

Resolution of quantitative traits into Mendelian factors by using a complete RFLP linkage map. Nature 335:721–726

Paterson AH, Damon S, Herwitt 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

Soller M, Brody T (1976) On the power of experimental designs for detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47:35–39

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