Mapping quantitative trait loci using four-way crosses

SHIZHONG XU

Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

(*Received 6 November 1995 and in revised form 25 April 1996*)

Summary

In plant species, typical gene mapping strategies use populations initiated from crosses between two inbred lines. However, schemes including more than two parents could be used. In this paper, a new approach is introduced which uses a four-way cross population derived from four inbred lines. The four-way cross design for mapping quantitative trait loci (QTLs) provides tests for QTL segregation in four lines simultaneously in one experiment. Therefore, it is a more economical strategy than one using line crosses between only two lines. The new strategy also increases the probability of detecting QTLs if they segregate in one line cross but not in the other. A multiple linear regression analysis is used for QTL detection. It is proven that the expected residual variance from the regression analysis differs from the pure environmental variance. Correction for the bias is proposed and verified by computer simulations.

1. Introduction

With the advent of new molecular technologies, a large number of polymorphic DNA markers can be generated with ease. Although most DNA markers are neutral to traits of economic importance, some markers may be tightly linked to loci that control the traits of interest so that they can be used to trace and locate the functional gene loci. This process is called mapping quantitative trait loci (QTLs). Statistical methods for QTL mapping are well developed in populations derived from an initial cross between two inbred lines. The typical schemes use a backcross, or the F_2 or more derived populations. Various statistical methods, typically maximum likelihood (Lander & Botstein, 1989; Jansen, 1993; Zeng, 1994) and least squares (Haley & Knott, 1992; Martinez & Curnow, 1992), are used for such line cross data.

In crosses involving two inbred lines, there are at most two alleles at any polymorphic locus. In most cases, the linkage phases between markers are known or can be easily inferred, and therefore the estimation of effects and locations of QTLs can be easily accomplished using a fixed linear model. Under such a model, the inference space is the two inbred lines, and therefore results from one line cross cannot be generalized to other line crosses. Because of this, if a position is tested and it turns out to be non-significant for the presence of a QTL, it does not mean that the genes at this locus have no function on the trait. The locus may contain genes functionally related to the trait but the gene effect cannot be detected because there is no segregation in the initial line cross, i.e. the two inbred lines are fixed for the same allele at that locus. If a different cross involving other inbred lines is chosen to initiate the mapping population, tests for the presence of a QTL at the very same locus may be significant. To expand the inference space and increase the probability of detecting QTLs, we should increase the number of inbred lines that are used to initiate the hybrid population (Rebai & Goffinet, 1993).

In this study, I investigated the use of a four-way cross population to detect QTLs. The four-way cross involves four homozygous lines (L_1, L_2, L_3 and L_4) and is analogous to an F_2 population but generated from hybridization between two different F_1 parents. The four-way cross, therefore, has a composition expressed as $(L_1 \times L_2) \times (L_3 \times L_4)$. Using similar notation, the composition of a conventional F_2 population (involving two inbred lines) can be expressed by $(L_1 \times L_2) \times (L_1 \times L_2)$, and a conventional backcross population by $(L_1 \times L_2) \times L_1$ or $(L_1 \times L_2) \times L_2$.

The four-way cross $(L_1 \times L_2) \times (L_3 \times L_4)$ is analogous to a two-way ANOVA experiment. It provides three tests simultaneously: one for QTL segregation between L_1 and L_2, one for segregation between L_3 and L_4, and one for the interaction (dominance) effects. In contrast, a backcross population, say $(L_1 \times L_2) \times L_1$, is analogous to a one-way ANOVA which can only test QTL segregation between L_1 and

https://doi.org/10.1017/S0016672300034066 Published online by Cambridge University Press

L_2. To test segregation between L_3 and L_4, another backcross, say $(L_3 \times L_4) \times L_3$, is required. The backcross design $(L_1 \times L_2) \times L_1$ is essentially a test for QTL segregation of $L_1 \times L_2$ using L_1 as a tester. The four-way cross $(L_1 \times L_2) \times (L_3 \times L_4)$, however, is a design that tests QTL segregation of $L_1 \times L_2$ using $L_3 \times L_4$ as a tester and, at the same time, tests segregation of $L_3 \times L_4$ using $L_1 \times L_2$ as a tester.

2. Methods

Let F_1^m denote the cross of $L_1 \times L_2$ and F_1^f the cross of $L_3 \times L_4$; the four-way cross offspring are derived from $F_1^f \times F_1^m$, i.e. $(L_3 \times L_4) \times (L_1 \times L_2)$. The superscripts m and f denote the male and female parental populations, respectively. It is not necessarily true that all parents from F_1^m population are males and all parents from F_1^f are females; the superscripts are used solely for distinguishing different populations and tracing the allelic origins of a QTL (Q) through markers (A and B).

Let $(A_1^m Q_1^m B_1^m)/(A_2^m Q_2^m B_2^m)$ and $(A_1^f Q_1^f B_1^f)/(A_2^f Q_2^f B_2^f)$ be the genotypes of F_1^m and F_1^f, respectively. In the offspring pool, there will be 16 possible genotype classes for the two flanking markers and four possible genotypes for the QTL, assuming that a marker allele in the male population is distinguishable from an allele in the female population. Let $p_{ik}^m = Pr(Q_i^m \mid M)$ and $p_{jk}^f = Pr(Q_j^f \mid M)$, be the corresponding allelic frequencies in the kth offspring conditional on flanking marker genotype (denoted by M). These conditional allelic frequencies (given in Table 1) are determined by r (the recombination fraction between loci A and B) and r_A or r_B (the recombination frequency of Q with A or B).

Since the QTL genotype is uncertain, the phenotype has a mixture of four distributions and must be described by a mixed model as shown below:

$$y_k = \mu + \alpha_i^m + \alpha_j^f + \delta_{ij} + \epsilon_k \quad \text{for} \quad i,j = 1,2, \qquad (1)$$

with a probability of $p_{ik}^m p_{jk}^f$, where y_k is the measurement of trait value for the kth plant ($k = 1, \ldots, n$), μ is the overall mean, α_i^m and α_j^f are the additive effects of alleles Q_i^m and Q_j^f, respectively, δ_{ij} is the dominance effect and ϵ_k is the error term with $N(0, \sigma_\epsilon^2)$. A formal treatment of such a mixed model is the maximum likelihood analysis. However, simple linear regression analysis can be used with virtually no loss in power

Table 1. *Conditional probability of QTL allele given flanking marker genotype*

Marker genotype [probability]	QTL allele			
	$p_{1k}^m = Pr(Q_1^m \mid M)$	$p_{2k}^m = Pr(Q_2^m \mid M)$*	$p_{1k}^f = Pr(Q_1^f \mid M)$	$p_{2k}^f = Pr(Q_2^f \mid M)$*
$A_1^m A_1^f B_1^m B_1^f$ $[\frac{1}{4}(1-r)^2]$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$
$A_1^m A_1^f B_1^m B_2^f$ $[\frac{1}{4}r(1-r)]$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$
$A_1^m A_2^f B_1^m B_1^f$ $[\frac{1}{4}r(1-r)]$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$
$A_1^m A_2^f B_1^m B_2^f$ $[\frac{1}{4}(1-r)^2]$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$
$A_1^m A_1^f B_2^m B_1^f$ $[\frac{1}{4}(1-r)]$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$
$A_1^m A_1^f B_2^m B_2^f$ $[\frac{1}{4}r^2]$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$
$A_1^m A_2^f B_2^m B_1^f$ $[\frac{1}{4}r^2]$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$
$A_1^m A_2^f B_2^m B_2^f$ $[\frac{1}{4}r(1-r)]$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$
$A_2^m A_1^f B_1^m B_1^f$ $[\frac{1}{4}r(1-r)]$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$
$A_2^m A_1^f B_1^m B_2^f$ $[\frac{1}{4}r^2]$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$
$A_2^m A_2^f B_1^m B_1^f$ $[\frac{1}{4}r^2]$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$
$A_2^m A_2^f B_1^m B_2^f$ $[\frac{1}{4}r(1-r)]$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$
$A_2^m A_1^f B_2^m B_1^f$ $[\frac{1}{4}(1-r)^2]$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$
$A_2^m A_1^f B_2^m B_2^f$ $[\frac{1}{4}r(1-r)]$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$	$(1-r_A) r_B/r$	$r_A(1-r_B)/r$
$A_2^m A_2^f B_2^m B_1^f$ $[\frac{1}{4}r(1-r)]$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$	$r_A(1-r_B)/r$	$(1-r_A) r_B/r$
$A_2^m A_2^f B_2^m B_2^f$ $[\frac{1}{4}(1-r)^2]$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$	$r_A r_B/(1-r)$	$(1-r_A)(1-r_B)/(1-r)$

* Note that $p_{2k}^s = 1 - p_{1k}^s$ for s = m, f.

https://doi.org/10.1017/S0016672300034066 Published online by Cambridge University Press

and efficiency (Haley & Knott, 1992). The simple regression method is accomplished by regressing y_k on the expected allelic and genotypic frequencies. Note that the expectation of y_k conditional on flanking marker genotype is

$$E(y_k\,|\,M) = \mu + \sum_{i=1}^{2} p_{ik}^{\mathrm{m}} \alpha_i^{\mathrm{m}} + \sum_{j=1}^{2} p_{jk}^{\mathrm{f}} \alpha_j^{\mathrm{f}} + \sum_{i=1}^{2}\sum_{j=1}^{2} p_{ik}^{\mathrm{m}} p_{jk}^{\mathrm{f}} \delta_{ij}. \tag{2}$$

There are nine unknown parameters in the model, but some of the coefficients of these parameters are linearly dependent. In fact, there are only four linearly independent coefficients to choose. Because $\Sigma_{i=1}^{2} p_{ik}^{\mathrm{m}} = 1$ and $\Sigma_{j=1}^{2} p_{jk}^{\mathrm{f}} = 1$, we have

$$p_{2k}^{\mathrm{m}} = 1 - p_{1k}^{\mathrm{m}} \quad \text{and} \quad p_{2k}^{\mathrm{f}} = 1 - p_{1k}^{\mathrm{f}}. \tag{3}$$

Substituting eqn (3) into (2) and doing some algebraic manipulations, we get

$$E(y_k\,|\,M) = \gamma + p_{1k}^{\mathrm{m}} \alpha^{\mathrm{m}} + p_{1k}^{\mathrm{f}} \alpha^{\mathrm{f}} + p_{1k}^{\mathrm{m}} p_{1k}^{\mathrm{f}} \delta, \tag{4}$$

where

$$\gamma = \mu + \alpha_2^{\mathrm{m}} + \alpha_2^{\mathrm{f}} + \delta_{22},$$

$$\alpha^{\mathrm{m}} = \alpha_1^{\mathrm{m}} - \alpha_2^{\mathrm{m}} + \delta_{12} - \delta_{22},$$

$$\alpha^{\mathrm{f}} = \alpha_1^{\mathrm{f}} - \alpha_2^{\mathrm{f}} + \delta_{21} - \delta_{22},$$

$$\delta = \delta_{11} - \delta_{12} - \delta_{21} + \delta_{22}.$$

Therefore, the model can be represented by

$$y_k = \gamma + p_{1k}^{\mathrm{m}} \alpha^{\mathrm{m}} + p_{1k}^{\mathrm{f}} \alpha^{\mathrm{f}} + p_{1k}^{\mathrm{m}} p_{1k}^{\mathrm{f}} \delta + e_k, \tag{5}$$

where e_k is a mixture of four normal distributions with a heterogeneous variance, but can be approximated by $N(0, \sigma_{\mathrm{e}}^2)$.

These newly defined parameters, $\{\gamma\alpha^{\mathrm{m}}\alpha^{\mathrm{f}}\delta\}$, are composite terms. γ is a rescaled mean, δ is the overall dominance effect, and α^{m} and α^{f} are the conditional average effects of an allelic substitution. The definition of the average effect of an allelic substitution is given by Falconer & Mackay (1995, pp. 112–114). The conditional average effect of an allelic substitution, say α^{m}, may be interpreted as follows: it is the difference between the average effect of individuals who received allele Q_1^{m} from their male parents and the average effect of individuals who received allele Q_2^{m} from their male parents, conditional on the gene received from the female parents having come from Q_2^{f}. This is because α_2^{f} has been included in γ. In other words,

$$\alpha^{\mathrm{m}} = G_{12} - G_{22} = (\alpha_1^{\mathrm{m}} + \alpha_2^{\mathrm{f}} + \delta_{12}) - (\alpha_2^{\mathrm{m}} + \alpha_2^{\mathrm{f}} + \delta_{22}).$$

Similarly,

$$\alpha^{\mathrm{f}} = G_{21} - G_{22} = (\alpha_2^{\mathrm{m}} + \alpha_1^{\mathrm{f}} + \delta_{21}) - (\alpha_2^{\mathrm{m}} + \alpha_2^{\mathrm{f}} + \delta_{22}),$$

where G_{ij} is the value of genotype $Q_i^{\mathrm{m}} Q_j^{\mathrm{f}}$.

Given the QTL position, the simple regression method generates unbiased estimates of $\{\gamma\alpha^{\mathrm{m}}\alpha^{\mathrm{f}}\delta\}$. The estimated residual variance, $\hat{\sigma}_{\mathrm{e}}^2$, is not an estimate of σ_ϵ^2 but of the environmental variance plus unexplained segregation at the QTL. Let us look first at $\mathrm{Var}(e_k)$, the variance of e_k. This variance is the conditional variance of y_k given the marker genotype, and it has the following form:

$$\mathrm{Var}(e_k) = \mathrm{Var}(y_k\,|\,M) = A_k(\alpha^{\mathrm{m}})^2 + B_k(\alpha^{\mathrm{f}})^2 + C_k\delta^2 + 2D_k(\alpha^{\mathrm{m}}\delta) + 2G_k(\alpha^{\mathrm{f}}\delta) + \sigma_\epsilon^2,$$

where $A_k = p_{1k}^{\mathrm{m}}(1 - p_{1k}^{\mathrm{m}})$, $B_k = p_{1k}^{\mathrm{f}}(1 - p_{1k}^{\mathrm{f}})$, $C_k = p_{1k}^{\mathrm{m}} p_{1k}^{\mathrm{f}}(1 - p_{1\mathrm{k}}^{\mathrm{m}} p_{1k}^{\mathrm{f}})$, $D_k = p_{1k}^{\mathrm{m}} p_{1k}^{\mathrm{f}}(1 - p_{1k}^{\mathrm{m}})$ and $G_k = p_{1k}^{\mathrm{m}} p_{1k}^{\mathrm{f}}(1 - p_{1k}^{\mathrm{f}})$. Therefore, the expected residual variance under model (5) is

$$\sigma_{\mathrm{e}}^2 = \bar{A}(\alpha^{\mathrm{m}})^2 + \bar{B}(\alpha^{\mathrm{f}})^2 + \bar{C}\delta^2 + 2\bar{D}(\alpha^{\mathrm{m}}\delta) + 2\bar{G}(\alpha^{\mathrm{f}}\delta) + \sigma_\epsilon^2, \tag{6}$$

where $\bar{A} = 1/n\Sigma_{k=1}^{n} A_k$, etc. Given the fact that $\sigma_{\mathrm{e}}^2 \neq \sigma_{\mathrm{e}}^2$ in general, one can adjust the regression estimate of σ_{e}^2 to obtain an estimate of the error variance σ_ϵ^2.

The standard F-test can be used for various hypothesis tests. The null hypothesis, $H_{0(\mathrm{m})}: \alpha^{\mathrm{m}} = 0$, is testing the QTL segregation in the $\mathrm{F}_1^{\mathrm{m}}$ population (the test is denoted by T_{m}), while $H_{0(\mathrm{f})}: \alpha^{\mathrm{f}} = 0$ is testing the QTL segregation in $\mathrm{F}_1^{\mathrm{f}}$ population (the test is denoted by T_{f}). The null hypothesis $H_{0(\mathrm{mf})}: \delta = 0$ tests the dominance effect (this test can be expressed by $T_{\mathrm{m}\times f}$). If the QTL location is treated as fixed (constant), under these hypotheses, the test statistic follows an F-distribution with degrees of freedom of 1 and $n-4$. An overall null hypothesis test for the segregation of a QTL in both $\mathrm{F}_1^{\mathrm{m}}$ and $\mathrm{F}_1^{\mathrm{f}}$ is $H_0: \alpha^{\mathrm{m}} = \alpha^{\mathrm{f}} = \delta = 0$ (this test is denoted by T). Under H_0, the test statistic for the overall test has an F-distribution with degrees of freedom of 3 and $n-4$.

To estimate the location of a QTL in a tested interval or chromosomal segment, one can examine the value of the test statistic for every position (with an increment of 1 or 2 cM). The estimated QTL location is the position showing the highest test statistic. With a variable QTL location, these test statistics may not follow F-distributions under the null hypotheses. Therefore, the critical values for the tests may not be found from the standard F-table. In practice, the permutation experiments of Churchill & Doerge (1994) may be used to find the correct critical values.

If there are multiple QTLs segregating within one chromosomal segment, interval mapping tends to be biased in terms of estimation of effects and locations of the QTLs. Composite interval mapping, however, can circumvent this problem by using other non-flanking markers simultaneously (Zeng, 1994). The composite interval mapping treats all markers outside the tested interval as QTLs and includes the indicator variables for other 'QTL' alleles in the model as covariates. The linear model is

$$y_k = \gamma + p_{1k}^{\mathrm{m}} \alpha^{\mathrm{m}} + p_{1k}^{\mathrm{f}} \alpha^{\mathrm{f}} + p_{1k}^{\mathrm{m}} p_{1k}^{\mathrm{f}} \delta + \sum_{t=1}^{q} (x_{tk}^{\mathrm{m}} \alpha_t^{\mathrm{m}} + x_{tk}^{\mathrm{f}} \alpha_t^{\mathrm{f}} + x_{tk}^{\mathrm{m}} x_{tk}^{\mathrm{f}} \delta_t) + e_k, \tag{7}$$

where q is the number of marker loci used as covariates, α_t^{m}, α_t^{f} and δ_t are corresponding effects for

https://doi.org/10.1017/S0016672300034066 Published online by Cambridge University Press

Table 2. *Indicator variables for marker alleles given marker genotype*

Marker genotype	x^{m}_{tk}	x^{f}_{tk}
$C^{\mathrm{m}}_1 C^{\mathrm{f}}_1$	1	1
$C^{\mathrm{m}}_1 C^{\mathrm{f}}_2$	1	0
$C^{\mathrm{m}}_2 C^{\mathrm{f}}_1$	0	1
$C^{\mathrm{m}}_2 C^{\mathrm{f}}_2$	0	0

the tth marker, and x^{m}_{tk} and x^{f}_{tk} are indicator variables taking values of 1 or 0, depending on the tth marker genotype. Let C denote the tth marker locus; the values of x^{m}_{tk} and x^{f}_{tk} are given in Table 2.

3. Simulation studies

The purposes of the simulations are to: (1) demonstrate the behaviour of the test statistic of an estimated genetic parameter; (2) investigate the estimation errors of the effect and location of a QTL using replicated experiments; (3) examine the statistical power of the QTL mapping procedure; and (4) verify the analytical relationship between the residual variance (σ^2_{ε}) in the regression model and the environmental variance (σ^2_{e}). The behaviour of composite interval mapping (using multiple markers) will not be investigated because it has previously been demonstrated and compared with simple interval mapping in the literature (e.g. Zeng, 1994; Jansen, 1994).

In the first experiment, one chromosomal segment was simulated with a length of 100 cM. The chromosome was covered by six markers separated by 20 cM intervals. One QTL located in the middle of the chromosome was simulated for a four-way cross population. It has been assumed that loci in each of the four inbred lines involved in the four-way cross had been alternatively fixed (i.e. four distinguishable alleles). Therefore, each allele has an equal representation in the four-way cross population. The parametric values were set at $\alpha^{\mathrm{m}}_1 = 10$, $\alpha^{\mathrm{f}}_1 = 10$, $\delta_{11} = 5$ and all other effects being equal to zero. Under such a model, the genetic values of the four genotypes are $G_{11} = 25$, $G_{12} = 10$, $G_{21} = 10$ and $G_{22} = 0$. Because the four genotypes have an equal frequency of 1/4, the mean is $(25+10+10+0)/4 = 11{\cdot}25$ and the genetic variance is $(25^2+10^2+10^2+0^2)/4 - 11{\cdot}25^2 = 79{\cdot}69 \approx 80$. I set two values for the pure environmental variance to calibrate the broad-sense heritability, h^2. In one situation, σ^2_{ε} was set at 80, which led to h^2 of approximately 50%. In another situation σ^2_{ε} was set at 240 so that h^2 was about 25%. No polygenic effect was simulated.

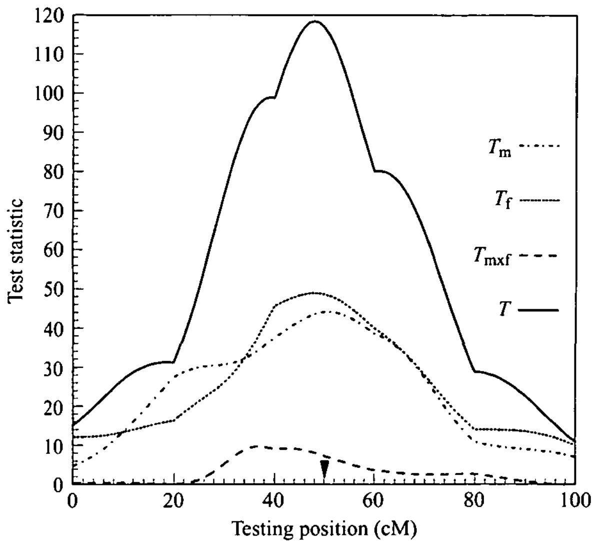

Fig. 1. A simulation example of QTL mapping from a four-way population. The test statistics are calculated and plotted every 1 cM. T_{m} is the test for QTL segregation in the male population ($\mathrm{F}^{\mathrm{m}}_1 = \mathrm{L}_1 \times \mathrm{L}_2$) and T_{f} is the test in the female population ($\mathrm{F}^{\mathrm{f}}_1 = \mathrm{L}_3 \times \mathrm{L}_4$). $T_{\mathrm{m\times f}}$ is the test for dominance effect and T is the overall test for the presence of a QTL. The arrow points to the true location of the QTL.

To evaluate the statistical power, one must first determine the critical values for the test statistics under the null hypothesis. Although the permutation tests of Churchill & Doerge (1994) are recommended, they are not feasible for simulation experiments. The critical values, however, can be approximated by repeated simulation experiments (Haley & Knott, 1992). To do this, I generated and mapped 1000 independent samples of the four-way cross population of size n (100, 250 and 500) under $H_0: \alpha^{\mathrm{m}} = \alpha^{\mathrm{f}} = \delta = 0$. The 95th percentile for the simulated test statistics over the 1000 samples was chosen as the critical value for the tests. The critical values seem to decrease a little as n increases; for example, when $n = 100$, 250 and 500, the critical values for T were 4·21, 4·02 and 3·99. Similar trends were also observed for the critical values of T_{m}, T_{f} and $T_{\mathrm{m\times f}}$. However, the observed increase was very small and may be due to chance because only 1000 samples were simulated. Nevertheless, the critical values of T_{m}, T_{f} and $T_{\mathrm{m\times f}}$ were about 7·3 in the three population sizes (100, 250 and 500). For simplicity, the empirical critical value for T was chosen to be 4·0. All other test statistics (T_{m}, T_{f} and $T_{\mathrm{m\times f}}$) took a common critical value of 7·3. Note that these critical values can only be used for the simulation experiment described above (six evenly distributed markers covering a chromosomal segment of 100 cM) and may not be used in general.

Fig. 1 shows the test statistic for each estimated genetic parameter plotted against the chromosomal position from a single simulated four-way cross population of size 500 with a broad-sense heritability of 0·50. It shows evidence of a QTL in the middle of the chromosome. The test statistic profiles (the curves) behave exactly as expected. In the middle of the chromosome, curves T_{m} and T_{f} have similar peaks

https://doi.org/10.1017/S0016672300034066 Published online by Cambridge University Press

Table 3. *Average estimates of QTL parameters and standard deviations over 100 replicated simulations (standard deviations are in parentheses)*

Heritability	Parameter	Parametric value	Estimate		
			$n = 100$	$n = 250$	$n = 500$
0·25	QTL location	50	49·29 (11·01)	49·55 (4·49)	50·02 (2·47)
	α^m	10	10·46 (5·47)	10·21 (3·39)	9·89 (2·36)
	α^f	10	10·24 (5·64)	10·29 (3·64)	9·83 (2·27)
	δ	5	3·27 (8·61)	4·63 (4·99)	4·88 (3·56)
	σ_e^2	256	249·66 (32·47)	255·33 (22·31)	253·77 (16·15)
0·50	QTL location	50	49·35 (4·57)	49·98 (2·34)	49·94 (1·34)
	α^m	10	9·68 (3·27)	9·99 (1·86)	10·02 (1·29)
	α^f	10	9·66 (3·19)	10·07 (1·76)	10·06 (1·46)
	δ	5	5·18 (4·87)	4·96 (2·94)	4·92 (1·94)
	σ_e^2	97	96·80 (14·20)	95·76 (8·30)	97·48 (5·86)

Table 4. *Average values of test statistics and empirical statistical powers obtained from 100 replicated simulation experiments*

		$n = 100$		$n = 250$		$n = 500$	
h^2	Test	Test statistic*	Power (%)	Test statistic	Power (%)	Test statistic	Power (%)
0·25	T	9·31 (3·88)	95	22·66 (6·63)	100	42·28 (7·79)	100
	T_m	5·21 (4·17)	25	10·58 (6·58)	63	18·55 (8·46)	95
	T_f	5·29 (4·75)	27	10·94 (7·19)	62	18·32 (8·83)	92
	$T_{m\times f}$	1·47 (1·91)	2	1·88 (2·31)	6	2·93 (2·96)	7
0·50	T	23·12 (7·65)	100	56·99 (10·73)	100	110·17 (15·08)	100
	T_m	9·94 (6·20)	57	24·62 (9·16)	99	47·04 (12·28)	100
	T_f	9·90 (6·50)	60	25·05 (8·88)	99	47·16 (13·06)	100
	$T_{m\times f}$	2·34 (3·49)	7	3·54 (3·67)	14	5·83 (4·65)	30

h^2 is the broad sense heritability.
* The standard deviations over 100 replicates are given in parentheses.

(because $\alpha^m = \alpha^f = 10$) and the curve $T_{m\times f}$ has a lower peak (because $\delta = 5$). The overall test statistic profile T has been meaningfully partitioned into three components. Replicated simulations with the same population size and heritability show similar patterns (data not shown).

In the following replicated simulation experiment, the population size was set at $n = 100$, 250 and 500 under each of the two heritability levels. Simulations were replicated 100 times in each parameter combination. The average estimates of α^m, α^f and δ, the location of the QTL and the residual variance (σ_e^2) are given in Table 3. According to eqn (6), the expected residual variances are 256 and 97, respectively, for $h^2 = 0·25$ and 0·50. The statistical powers under various situations are given in Table 4, with general trends behaving as anticipated.

It is interesting to see the results of simulations for a special case where there are two QTLs in the same chromosomal segment, one of which is segregating in F_1^m but not in F_1^f and the other segregating in F_1^f but not in F_1^m. In this simulation, I assumed that the first QTL is located at the 10 cM position and segregating in F_1^m but not in F_1^f, with parametric values of $\alpha_1^m = 10$, $\alpha_1^f = 0$ and $\delta_{11} = 0$. The second QTL was assumed to be located in the 70 cM position and segregating in F_1^f but not in F_1^m, with $\alpha_1^m = 0$, $\alpha_1^f = 10$ and $\delta_{11} = 0$. The environmental variance was set at $\sigma_e^2 = 80$. One hundred independent samples, each with 500 individuals, were simulated and mapped. The test statistic profiles of the 100 samples were averaged and then plotted in Fig. 2. The average curve for T shows two major peaks, indicating two QTLs, but each one of T_m and T_f shows one peak in the correct location. Since no dominance effect was simulated, the $T_{m\times f}$ curve is very low and flat.

4. Discussion

The proposed four-way cross QTL mapping strategy has a wider statistical inference space than the conventional single line cross methods. The latter are strictly appropriate for line crosses that are initiated from only two inbred lines. Invariably the two inbred lines are not a random sample from a larger population. Even if the lines are randomly sampled, given the limited number of lines involved, the statistical inference space cannot be extended to the

https://doi.org/10.1017/S0016672300034066 Published online by Cambridge University Press

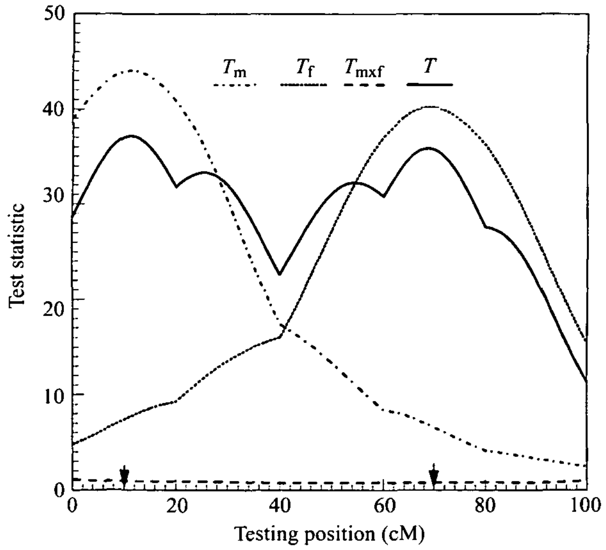

Fig. 2. Simulation of two QTLs in a four-way cross population. The QTL at the 10 cM position is present in the male parent ($F_1^m = L_1 \times L_2$) and the QTL at the 70 cM position is present in the female parent ($F_1^f = L_3 \times L_4$). Dominance effects are absent. The test statistic curves are calculated as the average of 100 replicated simulations. See legend to Fig. 1 for explanations of the test statistic curves.

larger population but is limited to the two lines selected. The four-way cross approach, however, can handle four lines at once, and thus the statistical inference space is the four lines instead of two. An experiment with a wider statistical inference space is preferable.

The four-way cross method can reduce or potentially eliminate the type II error caused by genetic drift. Type II error is defined as the probability that a QTL is not detected when in actuality it exists. There are two sampling steps in obtaining data for QTL analysis. The first step is to sample line crosses, and the second step is to sample individuals within a line cross. Accordingly, there are two kinds of type II errors associated with the two sampling processes. The first kind is caused by sampling lines and may be referred to as genetic 'drift error'. This kind of type II error can be interpreted as follows: if a particular locus contains a QTL that is segregating in the whole population, but, by chance, is not segregating in the sampled line crosses, then the QTL effect cannot be detected, no matter how good the statistical method is and how many individuals are sampled within the cross. The second kind of type II error is caused by a limited number of individuals being sampled within a cross. This second kind is what is normally called statistical 'sampling error'. Type II error of this kind can be diminished by using a more powerful statistical method and increasing the sample size of the mapped population. Simply by sampling one line cross, the single cross method has a maximum drift error, and this drift error has not been well documented in the literature.

The drift error can be shown in the following example. Suppose there are s alleles with equal frequency in a large population from which four inbred lines (representing four alleles) are randomly sampled. The probability that $L_1 \times L_2$ cross shows no segregation is $1/s = s(1/s)^2$, which is also the probability that $L_3 \times L_4$ shows no segregation. The probability that both crosses show no segregation is $1/s^2$. With a four-way cross experiment, if either $L_1 \times L_2$ or $L_3 \times L_4$ shows segregation at the locus of interest, the drift error is prevented. If the statistical sampling error is small, we may be able to detect the locus. Therefore, under the four-way cross design, the probability that the drift error is prevented is $1-1/s^2$. Under a backcross design, e.g. $(L_1 L_2) \times L_1$, if the allele sampled in L_1 is different from L_2, the drift error is prevented and this has a probability of $1-1/s$. Therefore, the probability that drift error is prevented in the four-way cross design is $(1-1/s^2)/(1-1/s) = (s+1)/s$ times as great as that in a backcross design. This ratio will be higher if the s alleles have different frequencies. Of course, the ratio tends to be unity as $s \to \infty$.

Assuming that the drift error has been prevented, the sampling error of the four-way cross design might be a little higher than that of the backcross design, but the difference is anticipated to be small. The reason is that in a backcross design L_1 is a tester for segregation of $L_1 \times L_2$ but L_1 is uniform, while in the four-way cross design $L_3 \times L_4$ serves as a tester for $L_1 \times L_2$ but $L_3 \times L_4$ is heterogeneous. As a result, one degree of freedom is lost when $L_1 \times L_2$ is tested. A smaller degree of freedom will cause a higher critical value for the test statistic and thus increase the type II error of the second kind. However, if the sample size is not too small, loss of one degree of freedom will have a negligible effect on the critical value. Therefore, in most situations, sampling errors of the two designs will be comparable. Let the sampling error be β for both designs, then the total type II error will be $\beta(1-1/s^2)$ in a four-way cross and $\beta(1-1/s)$ in a backcross. Therefore, by using a four-way cross design, the power increase relative to a backcross design will be $[1-\beta(1-1/s^2)]/[1-\beta(1-1/s)]$. The above argument is based on the assumption that the lines are randomly selected. In most line cross experiments lines are not randomly chosen; rather, only phenotypically very diverged lines are chosen. This will reduce the drift error in QTL analyses for the traits that served as the target of selection but may not affect other uncorrelated traits.

The interval mapping procedure in a four-way cross design described in this paper assumes that both parents (F_1^m and F_1^f) are heterozygotes and the status of an allele at any locus in F_1^m is different from that in F_1^f. In other words, for any interval bracketed by two markers, the flanking marker genotypes are $(A_1^m B_1^m)/(A_2^m B_2^m)$ for F_1^m and $(A_1^f B_1^f)/(A_2^f B_2^f)$ for F_1^f. In practice, this is rarely true. Sometimes it may be

https://doi.org/10.1017/S0016672300034066 Published online by Cambridge University Press

possible to infer the origins of flanking markers from the genotype at more distal markers that have the desired configuration. This implies that maps based on highly polymorphic markers such as microsatellites would be most useful with this design. A general solution for handling uninformative, partially informative and missing markers is to use several makers (multiple points) in the neighbourhood of the tested QTL to infer the origins of QTL alleles (Martinez & Curnow, 1994). Theoretically, this is not hard to do, but in practice it requires comprehensive computer programming. The multiple point approach to QTL mapping is very important in the four-way cross design and deserves further investigation.

I thank Timothy Close, Damian Gessler, Michael Roose, Richard Whitkus and two anonymous reviewers for helpful comments on and criticism of the manuscript. This research was supported by NRI Competitive Grants Program/USDA 95-37205-2313.

References

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Falconer, D. S. & Mackay, T. F. C. (1995). *Introduction to Quantitative Genetics*. New York: Longman.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Jansen, R. C. (1994). Controlling the Type I and Type II errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Martinez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

Martinez, O. & Curnow, R. N. (1994). Missing markers when estimating quantitative trait loci using regression mapping. *Heredity* **73**, 198–206.

Rebai, A. & Goffinet, B. (1993). Power of tests for QTL detection using replicated progenies derived from a diallele cross. *Theoretical and Applied Genetics* **86**, 1014–1022.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

https://doi.org/10.1017/S0016672300034066 Published online by Cambridge University Press