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 F2 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 (L1, L2, L3 and L4) and is analogous to an F4 population but generated from hybridization between two different F1 parents. The four-way cross, therefore, has a composition expressed as \((L_1 \times L_4) \times (L_2 \times L_3)\). Using similar notation, the composition of a conventional F2 population (involving two inbred lines) can be expressed by \((L_1 \times L_2) \times (L_1 \times L_2)\).

The four-way cross \((L_1 \times L_4) \times (L_2 \times L_3)\) is analogous to a two-way ANOVA experiment. It provides three tests simultaneously: one for QTL segregation between L1 and L2, one for segregation between L3 and L4, and one for the interaction (dominance) effects. In contrast, a backcross population, say \((L_1 \times L_3) \times L_4\), is analogous to a one-way ANOVA which can only test QTL segregation between L1 and...
L2. To test segregation between L3 and L4, another backcross, say (L3 x L4) x L3, is required. The backcross design (L1 x L2) x Lj is essentially a test for QTL segregation of L1 x L2 using L1 as a tester. The four-way cross (L1 x L2) x (L3 x L4), however, is a design that tests QTL segregation of L1 x L2 using L3 x L4 as a tester and, at the same time, tests segregation of L3 x L4 using L1 x L2 as a tester.

2. Methods

Let Fm denote the cross of L1 x L2 and Fi the cross of L3 x L4; the four-way cross offspring are derived from Fm x Fi, i.e. (L3 x L4) x (L1 x L2). The superscripts m and f denote the male and female parental populations, respectively. It is not necessarily true that all parents from Fm population are males and all parents from Fi 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 \( (A^m Q^m B^m)/(A^f Q^f B^f) \) and \( (A^i Q^i B^i)/(A^j Q^j B^j) \) be the genotypes of Fm and Fi, 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 \( p_{ik}^m = Pr(Q^m_i | M) \) and \( p_{ik}^f = Pr(Q^f_i | M) \) 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:

\[
y_k = \mu + \alpha^m_i + \alpha^f_i + \delta_i + \epsilon_k \quad \text{for} \quad i, j = 1, 2,
\]

with a probability of \( p_{ik}^m p_{ik}^f \), where \( y_k \) is the measurement of trait value for the kth plant \( (k = 1, \ldots, n) \), \( \mu \) is the overall mean, \( \alpha^m_i \) and \( \alpha^f_i \) are the additive effects of alleles \( Q^m_i \) and \( Q^f_i \), respectively, \( \delta_i \) is the dominance effect and \( \epsilon_k \) is the error term with \( \mathcal{N}(0, \sigma^2) \). 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 | \( p_{ik}^m = Pr(Q^m_i | M) \) | \( p_{ik}^f = Pr(Q^f_i | M) \) | \( p_{ik}^m = Pr(Q^m_i | M) \) | \( p_{ik}^f = Pr(Q^f_i | M) \) |
|-----------------------------|-----------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| \( A^m A^f B^m B^f \) \[\{1-(1-r)^2\] | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) |
| \( A^m A^f B^m B^f \) \[\{1-(1-r)^2\] | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) |
| \( A^m A^f B^m B^f \) \[\{1-(1-r)^2\] | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) |
| \( A^m A^f B^m B^f \) \[\{1-(1-r)^2\] | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) |
| \( A^m A^f B^m B^f \) \[\{1-(1-r)^2\] | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) |
| \( A^m A^f B^m B^f \) \[\{1-(1-r)^2\] | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) | \( (1-r_A)(1-r_B)/(1-r) \) | \( r_A r_B/(1-r) \) |

*Note that \( p_{ik}^m = 1 - p_{ik}^f \) for \( s = m, f \).*
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

\[
E(y_k | M) = \mu + \sum_{i=1}^{2} \beta_k^i \alpha^i + \sum_{j=1}^{2} \beta_k^j \alpha^j + \sum_{i=1,j=1}^{2} \beta_k^{ij} \alpha^i \alpha^j,
\]

(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 \( \Sigma_k \beta_k^i = 1 \) and \( \Sigma_k \beta_k^j = 1 \), we have

\[
p_k^+ = 1 - p_k^- \quad \text{and} \quad p_k^- = 1 - p_k^+.
\]

(3)

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

\[
E(y_k | M) = \gamma + p_k^+ \alpha^m + p_k^- \alpha^f + p_k^+ p_k^- \delta + e_k,
\]

(4)

where

\[
\gamma = \mu + \alpha^m + \alpha^f + \delta_{22},
\]

\[
\alpha^m = \alpha^m_1 - 2\alpha^m_2 - \delta_{12} - \delta_{22},
\]

\[
\alpha^f = \alpha^f_1 - 2\alpha^f_2 - \delta_{12} - \delta_{22},
\]

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

Therefore, the model can be represented by

\[
y_k = \gamma + p_k^+ \alpha^m + p_k^- \alpha^f + p_k^+ p_k^- \delta + e_k,
\]

(5)

where \( e_k \) is a mixture of four normal distributions with a heterogeneous variance, but can be approximated by \( N(0, \sigma^2) \).

These newly defined parameters, \( \gamma \alpha^m \alpha^f \delta \), are composite terms. \( \gamma \) is a rescaled mean, \( \delta \) is the overall dominance effect, and \( \alpha^m \) and \( \alpha^f \) 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 \( \alpha^m \), may be interpreted as follows: it is the difference between the average effect of individuals who received allele \( Q^m_2 \) from their male parents and the average effect of individuals who received allele \( Q^m_1 \) from their male parents, conditional on the gene received from the female parents having come from \( Q^m_1 \). This is because \( \alpha^m_2 \) has been included in \( \gamma \). In other words,

\[
\alpha^m = G_{12} - G_{22} = (\alpha^m_1 + \alpha^f_2 + \delta_{12}) - (\alpha^m_2 + \alpha^f_2 + \delta_{22}).
\]

Similarly,

\[
\alpha^f = G_{12} - G_{22} = (\alpha^m_1 + \alpha^f_2 + \delta_{12}) - (\alpha^m_2 + \alpha^f_2 + \delta_{22}),
\]

where \( G_{ij} \) is the value of genotype \( Q^m_i Q^f_j \).

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

\[
\text{Var}(e_k) = \text{Var}(y_k | M) = A_k(\alpha^m)^2 + B_k(\alpha^f)^2 + C_k \delta^2 + 2D_k(\alpha^m \delta) + 2G_k(\alpha^f \delta) + \sigma^2_e,
\]

where

\[
A_k = p_k^m(1 - p_k^-), \quad B_k = p_k^m(1 - p_k^+), \quad C_k = p_k^- p_k^+(1 - p_k^-), \quad D_k = p_k^- p_k^+(1 - p_k^+), \quad \text{and} \quad G_k = p_k^- p_k^+(1 - p_k^-). \quad \text{Therefore, the expected residual variance under model (5) is}
\]

\[
\sigma^2_e = \bar{A}(\alpha^m)^2 + B(\alpha^f)^2 + C\delta^2 + 2D(\alpha^m \delta) + 2G(\alpha^f \delta) + \sigma^2_e,
\]

(6)

where \( \bar{A} = 1/n \Sigma_{j=1}^{n} A_j \), etc. Given the fact that \( \sigma^2_{e1} + \sigma^2_{e2} \) in general, one can adjust the regression estimate of \( \sigma^2_e \) to obtain an estimate of the error variance \( \sigma^2_e \).

The standard \( F \)-test can be used for various hypothesis tests. The null hypothesis, \( H_{0mn} : \alpha^m = 0 \), is testing the QTL segregation in the \( F^m_1 \) population (the test is denoted by \( T_m \)), while \( H_{0nf} : \alpha^f = 0 \) is testing the QTL segregation in the \( F^f_1 \) population (the test is denoted by \( T_f \)). The null hypothesis \( H_{0mn} : \delta = 0 \) tests the dominance effect (this test can be expressed by \( T_{mn} \)).

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 \( F^m_1 \) and \( F^f_1 \) is \( H : \alpha^m = \alpha^f = \delta = 0 \) (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 increase 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

\[
y_k = \gamma + p_k^+ \alpha^m + p_k^- \alpha^f + p_k^+ p_k^- \delta + \sum_{i=1}^{q} (x_k^m \alpha^m + x_k^f \alpha^f + x_k^m x_k^f \delta) + e_k,
\]

(7)

where \( q \) is the number of marker loci used as covariates, \( \alpha^m, \alpha^f \) and \( \delta \) are corresponding effects for
the $t^{th}$ marker, and $x^m_{ik}$ and $x^f_{ik}$ are indicator variables taking values of 1 or 0, depending on the $t^{th}$ marker genotype. Let $C$ denote the $t^{th}$ marker locus; the values of $x^m_{ik}$ and $x^f_{ik}$ are given in Table 2.

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

<table>
<thead>
<tr>
<th>Marker genotype</th>
<th>$x^m_{ik}$</th>
<th>$x^f_{ik}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C^m C^f$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$C^m C^m$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C^f C^f$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C^f C^m$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 estimation errors of the effect and location of a QTL estimated genetic parameter; (2) investigate 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 each genotype from the four-way cross population of size $n$ (100, 250 and 500) under $H_0: \alpha^m = \alpha^f = \delta = 0$. 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 by one 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 $T_{max}$. 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 $T_{max}$ 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 40. All other test statistics ($T_m$, $T_f$ and $T_{max}$) 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. 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 ($F_{m} = L_1 \times L_2$) and $T_f$ is the test in the female population ($F_{f} = L_3 \times L_4$). $T_{max}$ 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.](https://www.cambridge.org/core/terms)
Table 3. Average estimates of QTL parameters and standard deviations over 100 replicated simulations (standard deviations are in parentheses)

<table>
<thead>
<tr>
<th>Heritability</th>
<th>Parameter</th>
<th>Parametric value</th>
<th>Estimate $n=100$</th>
<th>Estimate $n=250$</th>
<th>Estimate $n=500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>QTL location</td>
<td>50</td>
<td>49-29 (11-01)</td>
<td>49-55 (4-49)</td>
<td>50-02 (2-47)</td>
</tr>
<tr>
<td></td>
<td>$\alpha_m$</td>
<td>10</td>
<td>10-46 (5-47)</td>
<td>10-21 (3-39)</td>
<td>9-89 (2-36)</td>
</tr>
<tr>
<td></td>
<td>$\alpha'$</td>
<td>10</td>
<td>10-24 (5-64)</td>
<td>10-29 (3-64)</td>
<td>9-83 (2-27)</td>
</tr>
<tr>
<td></td>
<td>$\delta$</td>
<td>5</td>
<td>3-27 (8-61)</td>
<td>4-63 (4-99)</td>
<td>4-88 (3-56)</td>
</tr>
<tr>
<td>0.50</td>
<td>QTL location</td>
<td>50</td>
<td>49-35 (4-57)</td>
<td>49-98 (2-34)</td>
<td>49-94 (3-34)</td>
</tr>
<tr>
<td></td>
<td>$\alpha_m$</td>
<td>10</td>
<td>9-66 (3-27)</td>
<td>9-99 (1-86)</td>
<td>10-02 (1-29)</td>
</tr>
<tr>
<td></td>
<td>$\alpha'$</td>
<td>10</td>
<td>9-66 (3-19)</td>
<td>10-07 (1-76)</td>
<td>10-06 (1-46)</td>
</tr>
<tr>
<td></td>
<td>$\delta$</td>
<td>5</td>
<td>5-18 (4-87)</td>
<td>4-96 (2-94)</td>
<td>4-92 (1-94)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$h^2$</th>
<th>Test</th>
<th>Test statistic*</th>
<th>Power (%)</th>
<th>Test statistic</th>
<th>Power (%)</th>
<th>Test statistic</th>
<th>Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>$T$</td>
<td>9-31 (3-88)</td>
<td>95</td>
<td>22-66 (6-63)</td>
<td>100</td>
<td>42-28 (7-79)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$T_m$</td>
<td>5-21 (4-17)</td>
<td>25</td>
<td>10-58 (6-58)</td>
<td>63</td>
<td>18-55 (8-46)</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>$T_t$</td>
<td>5-29 (4-75)</td>
<td>27</td>
<td>10-94 (7-19)</td>
<td>62</td>
<td>18-32 (8-83)</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>$T_{mx}$</td>
<td>1-47 (1-91)</td>
<td>2</td>
<td>1-88 (2-31)</td>
<td>6</td>
<td>2-93 (2-96)</td>
<td>7</td>
</tr>
<tr>
<td>0.50</td>
<td>$T$</td>
<td>23-12 (7-65)</td>
<td>100</td>
<td>56-99 (10-73)</td>
<td>100</td>
<td>110-17 (15-08)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$T_m$</td>
<td>9-94 (6-20)</td>
<td>57</td>
<td>24-62 (9-16)</td>
<td>99</td>
<td>47-04 (12-28)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$T_t$</td>
<td>9-90 (6-50)</td>
<td>60</td>
<td>25-05 (8-88)</td>
<td>99</td>
<td>47-16 (13-06)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$T_{mx}$</td>
<td>2-34 (3-49)</td>
<td>7</td>
<td>3-54 (3-67)</td>
<td>14</td>
<td>5-83 (4-65)</td>
<td>30</td>
</tr>
</tbody>
</table>

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

(because $\alpha_m = \alpha' = 10$) and the curve $T_{mx}$ 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 $\alpha_m$, $\alpha'$ and $\delta$, the location of the QTL and the residual variance ($\sigma^2_{x}$) 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 $F_1$ but not in $F_2$ and the other segregating in $F_2$ but not in $F_1$. In this simulation, I assumed that the first QTL is located at the 10 cM position and segregating in $F_2$ but not in $F_1$, with parametric values of $\alpha_m = 10$, $\alpha' = 0$ and $\delta = 0$. The second QTL was assumed to be located in the 70 cM position and segregating in $F_1$ but not in $F_2$, with $\alpha_m = 0$, $\alpha' = 10$ and $\delta = 0$. The environmental variance was set at $\sigma^2_{x} = 80$. 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 them can only be located in the 10 cM position and segregating in $F_2$ or in $F_1$, but not in both. The statistical inference space cannot be extended to 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 conventional single line cross methods. The latter are strictly appropriate for line crosses that are initiated from only two inbred lines.

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 conventional single line cross methods. The latter are strictly appropriate for line crosses that are initiated from only two inbred lines.
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 $L_1 \times L_4$ cross shows no segregation is $1/s = x(1/s)^2$, which is also the probability that $L_2 \times L_4$ 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. $(L_1L_2) \times L_3$, if the allele sampled in $L_1$ is different from that in $L_3$, 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 $\beta$ 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 ($F^n_1$ and $F^n_1$) are heterozygotes and the status of an allele at any locus in $F^n_1$ is different from that in $F^n_1$. In other words, for any interval bracketed by two markers, the flanking marker genotypes are $(A^n_1B^n_2)/(A^n_2B^n_2)$ for $F^n_1$ and $(A^n_2B^n_1)/(A^n_1B^n_1)$ for $F^n_1$. In practice, this is rarely true. Sometimes it may be

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 ($F^n_1 = L_1 \times L_2$) and the QTL at the 70 cM position is present in the female parent ($F^n_1 = L_3 \times L_4$). 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.
Mapping QTLs using four-way crosses

It is 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


