Theor Appl Genet (1999) 98: 1014–1021

C. Xie · S. Xu

Mapping quantitative trait loci with dominant markers in four-way crosses

Received: 16 June 1998 / Accepted: 29 September 1998

Abstract A common problem in mapping quantitative trait loci (QTLs) is that marker data are often incomplete. This includes missing data, dominant markers, and partially informative markers, arising in outbred populations. Here we briefly present an iteratively reweighted least square method (IRWLS) to incorporate dominant and missing markers for mapping QTLs in four-way crosses under a heterogeneous variance model. The algorithm uses information from all markers in a linkage group to infer the QTL genotype. Monte Carlo simulations indicate that with half dominant markers, QTL detection is almost as efficient as with all co-dominant markers. However, the precision of the estimated QTL parameters generally decreases as more markers become missing or dominant. Notable differences are observed on the standard deviation of the estimated QTL position for varying levels of marker information content. The method is relatively simple so that more complex models including multiple QTLs or fixed effects can be fitted. Finally, the method can be readily extended to QTL mapping in full-sib families.

Key words Dominant markers · Four-way crosses · Missing data · QTL mapping · Weighted regression

Introduction

Dominant DNA markers are commonly used for mapping analysis in plants and animals due to the advent of efficient PCR (polymerase chain reaction)-based methods for amplifying random DNA sequences. Random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) are two such examples. Because little prior genomic information is required, the abundance and accessibility of PCR-based markers for virtually any species is a major advantage.

For mapping quantitative trait loci (QTLs), marker data are often incomplete. In addition to missing-data, another type of missing-marker data arises when markers are dominant, or when partially informative markers are used in experiments with outbred species. Dominant markers such as RAPDs generally show only two patterns, presence or absence of a band; consequently, a heterozygote has the same phenotype (band pattern) as one of the homozygotes. To deal with partial or missing-marker data, Lander and Green (1987) (cf. Kruglyak and Lander 1995) proposed a hidden Markov model to recover missing information. Jiang and Zeng (1997) used a similar idea for QTL mapping in line-cross experiments. Jansen (1996) developed a Monte Carlo EM algorithm via Gibbs sampling to handle missing and dominant markers. Although the algorithms developed by Lander and Green (1987) and by Jansen (1996) are general and may be used for outcrossing populations and complex pedigrees, these methods require extensive computation, particularly when the number of markers with incomplete information is large (Jiang and Zeng 1997). Furthermore, the original method of Lander and Green (1987) is feasible only for small pedigrees in human genetic studies.

The problems of dominant markers can be avoided by using backcross families, (doubled) haploid families, or recombinant inbred lines, but not with F_2, four-way crosses, or outbred populations (Plomion et al. 1996; Jiang and Zeng 1997). In view of this, we present an iteratively re-weighted least-square method (IRWLS) to incorporate dominant and missing markers in

Communicated by J. W. Snape

C. Xie (✉) · S. Xu
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA
Fax: +1 909-787-4437
E-mail: cxie@evolution.ucr.edu

four-way crosses. In particular, we modify Jiang and Zeng's (1997) algorithm to determine the distribution of QTL genotypes given observed marker phenotypes in four-way crosses. We choose four-way crosses as an example because the method of QTL mapping in four-way crosses sheds light upon experiments with outbred populations (Xu 1996) and is similar to that in full-sib families (Knott et al. 1997) and in sib mating designs (Xie et al. 1998). In addition, the IRWLS has an advantage over simple regression (REG) with regard to computational speed and the merit of maximum likelihood (ML) in a consideration of the heterogeneous residual variance (Xu 1998). Because the speed of IRWLS is nearly the same as that of REG it can be used for large genomic scanning or for multiple-data analyses, such as permutation tests (Churchill and Doerge 1994) and bootstrap construction of confidence intervals (Visscher et al. 1996). The method is relatively simple so that more complex models, including multiple QTLs or fixed effects, can be fitted simultaneously to increase the power of QTL detection by reducing the residual variance (Haley and Knott 1992; Jansen 1993; Zeng 1994). In the present paper, we briefly describe the IRWLS incorporating dominant and/or missing markers for QTL mapping in four-way crosses under a heterogeneous residual variance model. We further explore the statistical power and QTL parameter estimation of the proposed algorithm via Monte Carlo simulations.

Statistical methods

Genetic model

A four-way cross consists of two single crosses and can be expressed as $(L_1 \times L_2) \times (L_3 \times L_4)$. Let $Q_1^m Q_2^m$ and $Q_1^f Q_2^f$ be the genotypes at the QTL of $F_1^m(L_1 \times L_2)$ and $F_1^f(L_3 \times L_4)$, respectively. Consider a four-way cross between $Q_1^m Q_2^m \times Q_1^f Q_2^f (F_1^m \times F_1^f)$, there are four possible genotypes in the offspring pool, i.e. $Q_1^m Q_1^f$, $Q_1^m Q_2^f$ $Q_2^m Q_1^f$ and $Q_2^m Q_2^f$. Because the QTL genotype is unobservable, the phenotype has a mixture of four distributions. Let y_k be the phenotypic value of the kth progeny from a four-way cross. Then, it can be described by the following mixture model:

$$y_k = \mu + X_{k11}G_{11} + X_{k12}G_{12} + X_{k21}G_{21} + X_{k22}G_{22} + \varepsilon_k, \qquad (1)$$

where μ is the population mean, G_{ij} is the genetic effect of genotype, $Q_i^m Q_j^f$, ε_k is the residual error distributed as $N(0, \sigma_\varepsilon^2)$. The independent variables $\boldsymbol{X}_k = [X_{k11}, X_{k12}, X_{k21}, X_{k22}]$ are indicators of the four possible genotypes at the QTL and are defined as

$$X_{kij} = \begin{cases} 1 & \text{if } Q_i^m Q_j^f \text{ is true} \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i, j = 1, 2.$$

The genetic effect, G_{ij}, is a composite term. It consists of the additive effects of alleles from both paternal and maternal parents and the interaction effect (dominance) between alleles, i.e.,

$$\begin{bmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{bmatrix} = \begin{bmatrix} \alpha_1^m + \alpha_1^f + \delta_{11} \\ \alpha_1^m + \alpha_2^f + \delta_{12} \\ \alpha_2^m + \alpha_1^f + \delta_{21} \\ \alpha_2^m + \alpha_2^f + \delta_{22} \end{bmatrix},$$

where α_1^m and α_2^m are respectively the effects of the paternal and the maternal alleles of the male parent, α_1^f and α_2^f are respectively the effects of the paternal and the maternal alleles of the female parent, and δ_{ij} is the dominance effect.

Model (1) is over-parameterized because the sum of $\boldsymbol{X}_k$ equals the coefficient of μ. The equivalent form of model (1) can be rewritten as

$$y_k = X_{k11}\beta_1 + X_{k12}\beta_2 + X_{k21}\beta_3 + X_{k22}\beta_4 + \varepsilon_k, \qquad (2)$$

where $\beta_1 = G_{11} - \mu$, $\beta_2 = G_{12} - \mu$, $\beta_3 = G_{21} - \mu$, and $\beta_4 = G_{22} - \mu$.

Weighted regression analysis

Because the QTL genotype is unobservable, $\boldsymbol{X}_k$ are missing. However, the distribution of the QTL genotype can be inferred from linked markers, i.e., $\boldsymbol{X}_k$ can be estimated from marker genotypes. Let p_{kij} be the probability of the QTL genotype conditional on marker genotypes when $X_{kij} = 1$. Then, $p_{kij} = E(X_{kij} = 1 | I_M) = \Pr(X_{kij} = 1 | I_M)$, where I_M means the information of markers and an explicit definition is given later. By substituting the conditional expectation of X_{kij} into (2), we have

$$y_k = p_{k11}\beta_1 + p_{k12}\beta_2 + p_{k21}\beta_3 + p_{k22}\beta_4 + e_k. \qquad (3)$$

Note that the residual error e_k now is different from that given in equation (2).

In matrix notation, equation (3) can be written as

$$\mathbf{y} = \mathbf{P}\boldsymbol{\beta} + \mathbf{e}, \qquad (4)$$

where $\mathbf{P} = \{p_{kij}\}$ is an $N \times 4$ coefficient matrix, $\boldsymbol{\beta}$ is a 4×1 vector of unknown parameters and N is the number of individuals.

Now model (3) is a mixture of four distributions with a heterogeneous residual variance. The simple regression analysis under the assumption of a homogeneous variance provides a residual variance estimate that contains part of the QTL variance due to the uncertainty of QTL genotypes (Xu 1995, 1998).

With the weighted least-square method, the distribution of $\boldsymbol{X}_k$ is not required. Only the expectation and co-variance of $\boldsymbol{X}_k$ are needed. The expectation and covariance of $\boldsymbol{X}_k$ are:

$$E(\boldsymbol{X}_k) = [p_{k11}\ p_{k12}\ p_{k21}\ p_{k22}]$$

and

$$Var(\boldsymbol{X}_k) = \sum_k =$$

$$\begin{bmatrix} p_{k11}(1-p_{k11}) & -p_{k11}p_{k12} & -p_{k11}p_{k21} & -p_{k11}p_{k22} \\ -p_{k12}p_{k11} & p_{k12}(1-p_{k12}) & -p_{k12}p_{k21} & -p_{k12}p_{k22} \\ -p_{k21}p_{k11} & -p_{k21}p_{k12} & p_{k21}(1-p_{k21}) & -p_{k21}p_{k22} \\ -p_{k22}p_{k11} & -p_{k22}p_{k12} & -p_{k22}p_{k21} & p_{k22}(1-p_{k22}) \end{bmatrix}.$$

Let $Var(\mathbf{e}) = \mathbf{V}$, where $\mathbf{V}$ is a diagonal matrix with the kth element on the diagonal being

$$Var(e_k) = \boldsymbol{\beta}^{\mathrm{T}} \sum_k \boldsymbol{\beta} + \boldsymbol{\sigma}_\varepsilon^2, \qquad (5)$$

where $^{\mathrm{T}}$ denotes the matrix or vector transposition. Note that the first part, $\boldsymbol{\beta}^{\mathrm{T}} \sum_k \boldsymbol{\beta}$, is the variance not explained due to the uncertainty of the QTL genotype, and the second term $\boldsymbol{\sigma}_\varepsilon^2$, is the pure error variance (Xu 1995, 1998). The generalized least squares involves minimizing $(\mathbf{y} - \mathbf{P}\boldsymbol{\beta})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{P}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. This leads to

$$\hat{\boldsymbol{\beta}} = (\mathbf{P}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{P})^{-1} \mathbf{P}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{y} \qquad (6)$$

and

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N-4}[\mathbf{y} - \mathbf{P}\hat{\boldsymbol{\beta}}]^{\mathrm{T}}\mathbf{V}^{-1}[\mathbf{y} - \mathbf{P}\hat{\boldsymbol{\beta}}]. \quad (7)$$

Because $\mathbf{V}$ is a function of the unknown parameters, it must be updated using the estimated $\boldsymbol{\beta}$ and iteration is required. The process is repeated several times until convergence (Xu 1998).

After solving for, $\hat{\boldsymbol{\beta}}$, three independent estimable effects can be computed through $[\hat{\alpha}^m\ \hat{\alpha}^f\ \hat{\delta}]^{\mathrm{T}} = \mathbf{H}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, where $\hat{\alpha}^m$ and $\hat{\alpha}^f$ are the average effects of gene substitution for the male and female parents, respectively (Falconer and Mackay 1996), $\hat{\delta}$ is the interaction between paternal and maternal alleles or called the dominance effect, and $\mathbf{H}^{\mathrm{T}}$ is a 3×4 matrix containing the coefficients of contrasts as will be defined later. The three genetic effects are defined as (1) $\alpha^m = \alpha_1^m - \alpha_2^m$; (2) $\alpha^f = \alpha_1^f - \alpha_2^f$; and (3) $\delta = \delta_{11} - \delta_{12} - \delta_{21} + \delta_{22}$. Finally, the total genetic variance is

$$\hat{\sigma}_G^2 = \hat{\sigma}_A^2 + \hat{\sigma}_\delta^2,$$

where $\hat{\sigma}_A^2 = \frac{1}{4}[(\hat{\alpha}^m)^2 + (\hat{\alpha}^f)^2]$ is the additive genetic variance and $\hat{\sigma}_\delta^2 = \frac{1}{16}\hat{\delta}^2$ is the dominance variance.

Tests of hypotheses

To test the hypothesis that no QTLs are segregating, i.e., H_0: $[\alpha^m\ \alpha^f\ \delta]^{\mathrm{T}} = \mathbf{0}$, which is equivalent to $\mathbf{H}^{\mathrm{T}}\boldsymbol{\beta} = 0$, we use an F-test statistic,

$$F = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{H}[\mathbf{H}^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{P})^{-1}\mathbf{H}]^{-1}\mathbf{H}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \quad (8)$$

where

$$\mathbf{H}^{\mathrm{T}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

The overall test statistic can be partitioned into three separate tests each with a single degree of freedom:

H_1: $\alpha^m = 0$, which is equivalent to $\mathbf{h}_1^{\mathrm{T}}\boldsymbol{\beta} = 0$, where $\mathbf{h}_1^{\mathrm{T}} = [\frac{1}{2}\ \frac{1}{2}\ -\frac{1}{2}\ -\frac{1}{2}]$,

H_2: $\alpha^f = 0$, which is equivalent to $\mathbf{h}_2^{\mathrm{T}}\boldsymbol{\beta} = 0$, where $\mathbf{h}_2^{\mathrm{T}} = [\frac{1}{2}\ -\frac{1}{2}\ -\frac{1}{2}\ \frac{1}{2}]$,

H_3: $\delta = 0$, which is equivalent to $\mathbf{h}_3^{\mathrm{T}}\boldsymbol{\beta} = 0$, where $\mathbf{h}_3^{\mathrm{T}} = [1\ -1\ -1\ 1]$.

The F-test statistic for the specific effect becomes

$$F_i = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{h}_i[\mathbf{h}_i^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{P})^{-1}\mathbf{h}_i]^{-1}\mathbf{h}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} \quad (9)$$

for $i = 1, 2, 3$.

Inferring the QTL genotype from dominant markers

Assume that there are S ordered markers $(M_1, \ldots, M_S)$ with known recombination fractions on the tested chromosome and a QTL is located between markers t and $t+1$ $(1 \le t \le S-1)$. Define M_t as an indicator of the four possible genotypes ($A_1^mA_1^f$, $A_1^mA_2^f$, $A_2^mA_1^f$, $A_2^mA_2^f$) at the tth marker locus. Similarly, define M_q as an indicator of the four possible genotypes ($Q_1^mQ_1^f$, $Q_1^mQ_2^f$, $Q_2^mQ_1^f$, $Q_2^mQ_2^f$) at the QTL. Note that M_q here is equivalent to X_{ij} defined earlier. For convenience, we assign a single digit to denote a specific genotype at the tth marker locus or at the QTL, i.e.,

$$M_t = \begin{cases} 1 & \text{if } A_1^mA_1^f \\ 2 & \text{if } A_1^mA_2^f \\ 3 & \text{if } A_2^mA_1^f \\ 4 & \text{if } A_2^mA_2^f \end{cases}, \quad \text{and} \quad M_q = \begin{cases} 1 & \text{if } Q_1^mQ_1^f \\ 2 & \text{if } Q_1^mQ_2^f \\ 3 & \text{if } Q_2^mQ_1^f \\ 4 & \text{if } Q_2^mQ_2^f \end{cases}.$$

Let $p_{tu} = \Pr(M_t = u|I_M)$ for $u = 1, \ldots, 4$ be the probability that an offspring has a given marker genotype, $M_t = u$, conditional on the genotypes of its parents at marker t. When a marker is fully informative, all four genotypes are distinguishable, that is, $p_{tu} = 1$ for the observed genotype $M_t = u$ and $p_{tu} = 0$ otherwise. When a marker is unobserved (or missing), $p_{tu} = 1/4$ for all four possible genotypes. When dominant or partially informative markers occur, the marker genotype is ambiguous. For example, if A_1 is dominant over A_2, a dominant phenotype could be either A_1A_1 or A_1A_2 with an equal probability of $p_{tu} = 1/2$.

The sequence of marker and QTL genotypes, $\{M_1 \ldots M_tM_qM_{t+1} \ldots M_S\}$, forms a Markov chain with transitions between states caused by recombination. Given the recombination fraction (r) between two markers, the transition matrix between two adjacent markers is

$$\mathbf{T} = \begin{bmatrix} P(M_{t+1}=1|M_t=1) & P(M_{t+1}=2|M_t=1) & P(M_{t+1}=3|M_t=1) & P(M_{t+1}=4|M_t=1) \\ P(M_{t+1}=1|M_t=2) & P(M_{t+1}=2|M_t=2) & P(M_{t+1}=3|M_t=2) & P(M_{t+1}=4|M_t=2) \\ P(M_{t+1}=1|M_t=3) & P(M_{t+1}=2|M_t=3) & P(M_{t+1}=3|M_t=3) & P(M_{t+1}=4|M_t=3) \\ P(M_{t+1}=1|M_t=4) & P(M_{t+1}=2|M_t=4) & P(M_{t+1}=3|M_t=4) & P(M_{t+1}=4|M_t=4 \end{bmatrix} = \begin{bmatrix} (1-r)^2 & r(1-r) & r(1-r) & r^2 \\ r(1-r) & (1-r)^2 & r^2 & r(1-r) \\ r(1-r) & r^2 & (1-r)^2 & r(1-r) \\ r^2 & r(1-r) & r(1-r) & (1-r)^2 \end{bmatrix}.$$

The transition probability matrix between the marker and the hypothesized QTL is similarly defined. Our purpose here is to estimate the conditional probability $\Pr(M_q = u\,|\,I_M)$ at the QTL from $(M_1 \ldots M_s)$. According to Bayes' theorem, the probability of the QTL genotype conditional on marker information (I_M) can be computed as:

$$\Pr(M_q = u|I_M) = \frac{\Pr(M_q = u)\Pr(I_M|M_q = u)}{\sum_{u=1}^{4}\Pr(M_q = u)\Pr(I_M|M_q = u)} = \frac{\Pr(I_M|M_q = u)}{\sum_{u=1}^{4}\Pr(I_M|M_q = u)}, \quad (10)$$

where $I_M = \{M_t\}$ are marker genotypes for $t = 1, \ldots, S$, and $\Pr(M_q = u) = 1/4$ is the prior probability of the QTL genotype. Note that $\Pr(M_q = u\,|\,I_M) = \Pr(X_{ij} = 1\,|\,I_M) = p_{ij}$ which is defined in equation (3). Since $\{M_1 \ldots M_tM_qM_{t+1} \ldots M_s\}$ is a Markov chain, $\Pr(I_M|M_q = u)$ can be computed by a hidden Markov model (Lander and Green 1987; Jiang and Zeng 1997; Xu and Gessler 1998). In matrix notation, we have

$$\Pr(I_M|M_q = u) = \mathbf{1}^{\mathrm{T}}\mathbf{D}_1\mathbf{T}_{12}\mathbf{D}_2 \ldots \mathbf{T}_{tq}\mathbf{D}_{qu}\mathbf{T}_{qt+1} \ldots \mathbf{D}_{S-1}\mathbf{T}_{S-1S}\mathbf{D}_S\mathbf{1}, \quad (11)$$

where

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{D}_t = \begin{bmatrix} p_{t1} & 0 & 0 & 0 \\ 0 & p_{t2} & 0 & 0 \\ 0 & 0 & p_{t3} & 0 \\ 0 & 0 & 0 & p_{t4} \end{bmatrix}$$

Table 1 Probabilities $p_{tu} = \Pr(M_t = u|I_M)$ conditional on the mating type and the observed offspring phenotype, assuming that the linkage phases in the parents are known (i.e., gene order) and that A is a dominant allele and O is a recessive allele

Mating type	Progeny type	p_{t1}	p_{t2}	p_{t3}	p_{t4}
$A_1A_2 \times A_1A_2$	A-	1/4	1/4	1/4	1/4
$A_1A_2 \times A_1O_2$	A-	1/4	1/4	1/4	1/4
$A_1A_2 \times O_1O_2$	A-	1/4	1/4	1/4	1/4
$A_1O_2 \times A_1O_2$	A-	1/3	1/3	1/3	0
	OO	0	0	0	1
$A_1O_2 \times O_1O_2$	A-	1/2	1/2	0	0
	OO	0	0	1/2	1/2
$O_1O_2 \times A_1O_2$	A-	1/2	0	1/2	0
	OO	0	1/2	0	1/2
$O_1O_2 \times O_1O_2$	OO	1/4	1/4	1/4	1/4

and

$$\mathbf{D}_{qu} = \begin{cases} \text{diag}[1\ 0\ 0\ 0] & \text{for } M_q = 1 \\ \text{diag}[0\ 1\ 0\ 0] & \text{for } M_q = 2 \\ \text{diag}[0\ 0\ 1\ 0] & \text{for } M_q = 3 \\ \text{diag}[0\ 0\ 0\ 1] & \text{for } M_q = 4 \end{cases}.$$

To determine the probability of $\Pr(I_M|M_q = u)$ using (1 1), the key is to compute the $\mathbf{D}_t$ matrix or $p_{tu} = \Pr(M_t = u|I_M)$ at each marker locus. With fully informative markers, $\mathbf{D}_t$ has one diagonal element of 1 for the observed marker genotype and 0 elsewhere. With dominant or partially informative markers, $\mathbf{D}_t$ may have more than one non-zero diagonal element. To simplify the problem, let us consider dominant markers such as RAPDs where they generally show only two patterns: presence or absence of a band. Define A as a dominant allele and O as a recessive allele. In a four-way cross, the linkage phases of marker loci in the parents are known. Furthermore, the AA and AO genotypes in the parents can be inferred from the grand-parents. However, unlike an F_2 family where there is only one mating type, a four-way cross may include three genotypes (AA AO OO) and nine mating types $(\text{AA AO OO})^2$. Computation of p_{tu} depends on the mating type as well as the offspring's marker genotypes. Table 1 presents p_{tu} values and possible mating types in a four-way cross. The mating types (e.g., $O_1A_2 \times O_1A_2$) differ from $A_1O_2 \times A_1O_2$ due to the fact that the linkage phases are ignored.

The algorithm described in this paper uses information from all markers in a linkage group simultaneously to determine the QTL genotype (Fulker et al. 1995; Kruglyak and Lander 1995; Xu and Gessler 1998).

Simulation studies

To evaluate the effects of dominant and missing markers on mapping QTLs, we simulated one chromosome of length 100 cM with 11 markers evenly spaced along the chromosome. One QTL was simulated at position 25 cM with a sample size of 300 individuals. The linkage map is assumed to be known. Five levels of marker information content were investigated: (1) all markers are co-dominant with no missing markers (standard); (2) 50% of the loci in the parents are randomly set to dominant with no missing markers in the offspring; (3) 50% of the loci in the offspring are randomly set to missing markers; (4) 50% of the loci in the parents are randomly set to dominant and 50% of the loci in the offspring are randomly set to missing; and (5) all loci are dominant.

In the simulation, up to four alleles for co-dominant markers and two alleles (presence or absence) for dominant markers at each locus in each of the two parents were sampled at random from the base population with an equal frequency for each allele. The variance of the environmental effect was set at $\sigma_e^2 = 1.0$. The interaction effect was set at $\delta = 0$. The average effect of a gene substitution was examined at three levels, $\alpha^m = \alpha^f = 0.324$, 0.594 and 1.155, which corresponded to the additive genetic variances of $\sigma_G^2 = \sigma_A^2 = 0.0526$, 0.1765 and 0.6670, or equivalently to $h^2 = 0.05$, 0.15 and 0.40, respectively. Under each condition, the simulation was repeated 120 times. The standard error of a parametric estimate is calculated from the standard deviation of the estimates among 120 replicates. The statistical power is determined by counting the number of runs out of 120 replicates which have the overall F-test statistics (Eq. 8) greater than an empirical threshold. To estimate the strength of a false positive signal, we ran an additional 1000 simulations with no QTL segregating. The empirical threshold under each condition was then obtained by determining the 95th percentile of the highest F-test statistic (Eq. 8) from the list of 1000 runs under the null model.

Results

The overall F-test statistics with a heritability of 0.15 for five levels of marker information content are plotted against the genomic position (Fig. 1). Results indicate that the highest level of marker information (1) with co-dominant and no missing markers produces the highest and the most-narrow peak. The test statistics decrease as the marker information content diminishes. The marker information (2) with 50% dominant and no missing markers has a greater test statistic than the marker information (3) with 50% missing markers. Similarly, the marker information (5) with 100% dominant and no missing markers has a test statistic larger than the marker information (4) with 50% dominant and 50% missing markers. The other important feature shown by the figure is that, with less marker information, the peak of the curve (test statistic) tends to be flat, indicating that QTL detection has a great uncertainty.

More detailed results on estimated QTL effect, position, QTL heritability and residual variance with varying marker information contents are presented in Table 2. The results indicate that precision of estimated QTL parameters generally decreases as more markers become missing or partially missing (dominant). There is little difference on the point estimates of QTL effects, heritability and residual variance for different marker information contents except for 100% dominant markers (5). However, under varying levels of marker information content notable differences are observed on the standard deviation of the estimated QTL position. In the case of all dominant markers, the estimated QTL effects are generally biased downward and the QTL position is biased toward the center. The levels of QTL heritability have a strong effect on the precision of the estimated QTL position but have a small effect on the estimated QTL effects and residual variance. As expected, a high QTL heritability

decreases the standard deviation of the estimated QTL position. When the QTL effect is small, as in the case of $h^2 = 0.05$, the estimated QTL position is biased. This bias is caused by some runs where the QTL effect is not significant. In these situations, the QTL position, on average, tends to be close to the center.

The empirical threshold values of test statistics over 1000 replicated simulations are reported in columns 6 and 7 of Table 3. Slight differences among these critical values across the levels of marker information content can be observed with the co-dominant and no missing markers having the highest critical values, whilst the dominant markers have the lowest critical values. Similar results were observed by Xu (1998) in F_2 families. The average test statistics and the power estimates (Type-I error rate at $\gamma = 0.05$ and $\gamma = 0.01$) over 120 replicated simulations are summarized in Table 3. Generally, the test statistics and the power of QTL detection increase as the marker information content increases. When the power is not already high, as in the case of $h^2 = 0.05$, marker information (1) with all co-dominant and no missing markers has the greatest power. In contrast, marker information (5) with all dominant markers has the lowest power. As expected, when the power is already high (e.g. $h_g^2 = 0.4$), the difference in the power between varying levels of marker information content disappears.

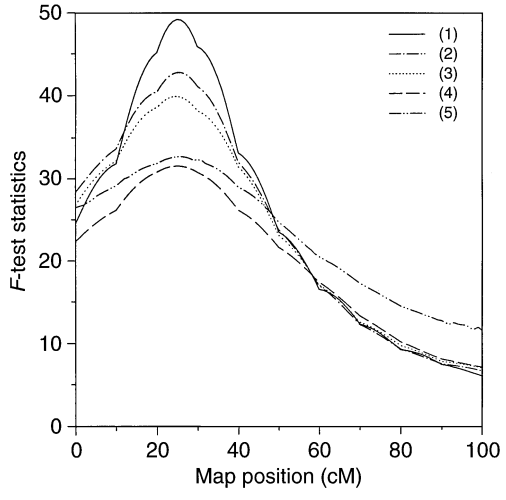

Fig. 1 Comparison of the *F*-test statistics of QTL mapping in a four-way cross population with a size of 300 individuals for five levels of marker information content: (1) co-dominant with no missing markers (standard); (2) 50% dominant; (3) 50% missing; (4) 50% dominant and 50% missing; (5) 100% dominant. Eleven markers are evenly spaced along a 100-cM chromosome. A single QTL accounting for 15% additive genetic variance is at the 25-cM position

A four-way cross is analogous to a two-way ANOVA experiment. The overall *F*-test statistic (T) can be partitioned to three separate tests: one (T_m) for QTL segregation in the male parent ($L_1 \times L_2$), one (T_f) for QTL segregation in the female parent ($L_3 \times L_4$), and one ($T_{m\times f}$) for the interaction (dominance) effects (Xu 1996). Figure 2 gives an example of the test statistics for a QTL with a heritability of 0.15 and located at 25 cM of the chromosome. The profiles of the test statistics (curves) perform exactly as expected. The overall test (T) for the presence of a QTL has a signal twice as great either T_m or T_f separately. Whereas T_m and T_f have similar curves because $\alpha^m = \alpha^f = 0.297$, and $T_{m\times f}$ has a flat curve near zero because $\delta = 0$.

Table 2 Results on estimated QTL effects, position, heritability and residual variance for five levels of marker information content. Parameters used for simulation are: $\delta = 0$, $\sigma_\varepsilon^2 = 1.0$ and QTL position (QP) = 25 cM. Standard deviations among 120 replicates are given in parentheses

α^m (α^f)	h^2	MI[a]	Estimate				
			$\frac{1}{2}(\hat{\alpha}^m + \hat{\alpha}^f)$	$\hat{\delta}$	$\hat{\text{QP}}$(cM)	$\hat{h}^2$	$\hat{\sigma}_\varepsilon^2$
0.324	0.05	(1)	0.318 (0.097)	−0.023 (0.326)	26.37 (16.64)	0.053 (0.028)	0.996 (0.081)
		(2)	0.314 (0.103)	−0.063 (0.370)	27.20 (15.23)	0.049 (0.028)	0.990 (0.086)
		(3)	0.330 (0.100)	0.033 (0.389)	28.98 (18.58)	0.056 (0.031)	0.988 (0.090)
		(4)	0.340 (0.148)	0.057 (0.572)	30.49 (21.80)	0.066 (0.051)	0.970 (0.092)
		(5)	0.319 (0.256)	0.043 (1.083)	31.72 (25.39)	0.080 (0.111)	0.888 (0.262)
0.594	0.15	(1)	0.603 (0.093)	−0.027 (0.324)	25.30 (4.35)	0.155 (0.044)	0.993 (0.088)
		(2)	0.594 (0.097)	0.024 (0.343)	25.35 (6.98)	0.151 (0.046)	0.993 (0.095)
		(3)	0.596 (0.116)	−0.014 (0.374)	24.40 (6.62)	0.156 (0.057)	0.980 (0.088)
		(4)	0.589 (0.117)	0.011 (0.434)	26.19 (12.52)	0.153 (0.061)	0.970 (0.108)
		(5)	0.540 (0.351)	0.120 (1.350)	28.24 (19.07)	0.140 (0.343)	0.928 (0.215)
1.155	0.40	(1)	1.153 (0.089)	0.011 (0.267)	24.90 (2.39)	0.400 (0.045)	0.993 (0.076)
		(2)	1.162 (0.115)	0.070 (0.341)	25.50 (3.63)	0.406 (0.064)	0.989 (0.107)
		(3)	1.173 (0.098)	0.019 (0.366)	25.08 (3.61)	0.411 (0.054)	0.984 (0.107)
		(4)	1.154 (0.136)	−0.040 (0.555)	25.71 (5.71)	0.400 (0.076)	0.990 (0.145)
		(5)	1.128 (0.563)	−0.175 (1.984)	26.90 (12.55)	0.391 (0.140)	0.931 (0.325)

[a] MI, marker information: (1) co-dominant with no missing markers (standard); (2) 50% dominant; (3) 50% missing; (4) 50% dominant and 50% missing; (5) 100% dominant

Table 3 Empirical threshold values, the overall *F*-test statistics and the power (%) of QTL detection for five levels of marker information content. Standard deviations among 120 replicates are given in parentheses

h^2	MI[a]	Test statistic	Power (%)		Threshold value	
			$\gamma = 0.05$[b]	$\gamma = 0.01$	95%	99%
0.05	(1)	18.01 (8.25)	68.33	44.17	12.70	17.25
	(2)	15.15 (6.81)	63.33	43.33	11.74	15.46
	(3)	15.31 (6.86)	60.83	30.00	12.52	17.25
	(4)	14.02 (7.55)	63.33	33.33	11.49	15.70
	(5)	13.62 (8.01)	62.50	28.33	10.85	16.77
0.15	(1)	50.80 (16.16)	100.00	100.00	12.71	17.40
	(2)	43.81 (14.31)	98.33	97.50	11.69	14.93
	(3)	41.86 (16.42)	98.33	97.50	12.52	15.54
	(4)	33.47 (12.93)	97.50	92.50	11.50	15.58
	(5)	31.08 (15.25)	87.50	82.50	11.53	15.34
0.40	(1)	168.84 (34.54)	100.00	100.00	12.53	15.92
	(2)	150.44 (40.63)	100.00	100.00	11.56	15.26
	(3)	134.21 (28.79)	100.00	100.00	12.22	15.80
	(4)	102.10 (35.54)	100.00	100.00	11.82	15.81
	(5)	100.70 (42.31)	98.33	98.33	11.20	14.05

[a] See Table 2 for notations of MI
[b] γ is the Type-I error rate

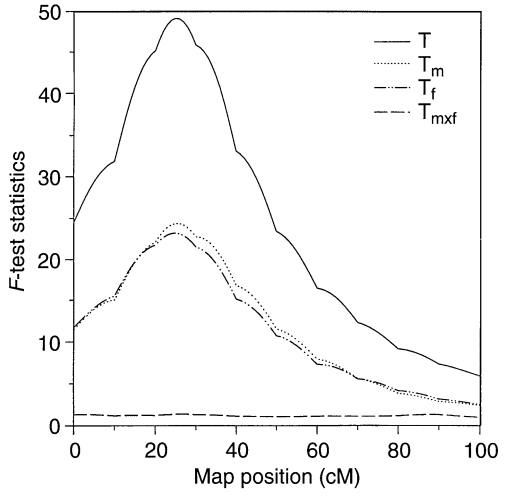

Fig. 2 Profile of the *F*-test statistics for QTL mapping in a four-way cross population with a size of 300 individuals. Eleven co-dominant markers are evenly spaced along a 100-cM chromosome. A single QTL explaining 15% additive genetic variance is at the 25-cM position. T is the overall test for the presence of a QTL; T_m is the test for QTL segregation in the male parent; T_f is the test for QTL segregation in the female parent; and $T_{m\times f}$ is the test for the QTL dominance effect

Discussion

Saturated maps have been constructed for a number of species using restricted fragment length polymorphisms (RFLPs), but RAPDs and AFLPs provide significant advantages in speed and efficiency for constructing such maps in order to locate important genes, especially in some allogamous species such as forest trees where little prior genetic information is available (Williams et al. 1993; Grattapaglia and Sederoff 1994; Plomion et al. 1996). Because PCR-based markers are rich, and less expensive, genetic mapping of genes (or QTLs) by using dominant markers is widely applicable and will become a common practice. In view of this, concerns have been expressed on the use of these markers in genetic mapping due to their partial information content (Martínez and Curnow 1994; Knapp et al. 1995; Jansen 1996; Plomion et al. 1996; Jiang and Zeng 1997).

In this paper, we demonstrate a method for QTL mapping with dominant or missing markers in four-way crosses. A four-way cross involves four parental lines. It differs from an F_2 family in that a four-way cross population may have a maximum of four alleles at each marker locus or at the QTL, whereas an F_2 family may have only two alleles at each locus. The four-way cross $(L_1 \times L_2) \times (L_3 \times L_4)$ design is also analogous to two backcrosses. This essentially increases the possibility of detecting QTLs compared to that of an F_2 family because the probability of no QTL segregation in both $L_1 \times L_2$ and $L_3 \times L_4$ is less than that in only one of them (Xu 1996). A full-sib family may also have four alleles at each locus. In this sense, the four-way cross is more similar to a full-sib family (Knott et al. 1997) and to sib mating designs (Xie et al. 1998). Given that the linkage phase in a full-sib family can be inferred, QTL mapping in a four-way cross population is identical to that in the full-sib family or to the population resulting from sib matings. Therefore, the method presented here can be readily extended to outbred populations.

Xu (1998) has shown that the weighted regression method (IRWLS) is virtually identical to the ML method in terms of computational results in F_2 families, yet it retains the advantages of REG in simplicity and high computational speed. In this paper, we

incorporate IRWLS with dominant or missing markers for QTL mapping in four-way crosses. Our results indicate that dominant markers are useful for QTL detection, but they are not as efficient as co-dominant markers in determining the QTL position and in estimating QTL parameters. The power of QTL detection with half dominant markers (2) is almost the same as that with all co-dominant markers (1) and is higher than that with half missing markers (3) when the power is not already high (e.g., $h^2 = 0.05$). However, the standard deviations of estimated QTL parameters are virtually equivalent in the levels of marker information content in (2) and (3). When QTL mapping uses all dominant markers, the estimated QTL position and parameters are severely biased and have exceptionally large standard deviations. This is caused by the large error in determining the QTL distribution due to the uncertainty of marker genotypes, whereas the latter results from the partial information content of dominant markers. In addition, we have found that the QTL parameters are inestimable in approximately 4% of the simulation runs with all dominant markers, because $\mathbf{V}^{-1}$ or $(\mathbf{P}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{P})^{-1}$ can not be inverted in these cases due to an insufficient marker information content.

The algorithm presented here can be used to fit for multiple QTLs that are segregating in the same linkage group. In practice, two types of multiple QTL models can be distinguished: composite interval mapping (Zeng 1994) and multiple QTL regression model (Haley and Knott 1992). In our case with dominant markers, the efficiency of composite interval mapping is not known because the flanking markers are not fully informative. However, the multiple QTL regression model is also applicable to the weighted regression method (IRWLS). With multiple QTLs, an additional term, $\mathbf{P}\boldsymbol{\beta}$, for each QTL is added into Model (4). Similarly, the residual variance can be partitioned to the corresponding term for each QTL due to the uncertainty of QTL genotypes and the pure error variance (cf. Eq. 5). Because of the simplicity of IRWLS, the stepwise method of regression can be used to search for multiple QTLs. However, the efficiency of the multiple QTL model requires further investigation.

When markers are fully informative, the position of the QTL is restricted within the bracket of the two nearest flanking markers if there is a QTL. With missing and/or dominant markers, QTL mapping extracts extra information beyond the two nearest flanking markers. Equivalently, the resulting effect can be seen as increasing the bracket (distance) of two flanking markers so that the QTL is more variable in location within the bracket. In our simulation, the linkage map is assumed to be known. However, in practice the linkage map is often less certain or requires to be determined from the same dataset that is used for QTL mapping. Jiang and Zeng (1997) have investigated the consequence of missing and dominant markers on marker linkage map construction in an F_2 population. Their results indicate that the proportion of correct linkage order decreases as more markers become missing or partially missing, and is low when markers are dominant and recessive in alternate order. Nevertheless, the proportion of intervals with correct flanking markers remains reasonably high. It would be expected that an ambiguous linkage map significantly affects QTL mapping, and particularly on QTL parameter estimates. Hence, it is important for QTL mapping to combine dominant and co-dominant markers. The significance of co-dominant markers lies in the fact that they provide landmarks in various positions of the genome.

We have developed a FORTRAN program to perform IRWLS analysis for QTL mapping with dominant or missing markers in four-way crosses. This program is available upon request.

Acknowledgments The authors thank Damian Gessler for his helpful comments. This research was supported by the National Institutes of Health Grant GM55321-01 and the USDA National Research Initiative Competitive Grants Program 97-35205-5075 to S.X.

References

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. 4th edn. Longman, England, Essex

Fulker DW, Cherny SS, Cardon LR (1995) Multipoint interval mapping of quantitative trait loci using sib pairs. Am J Hum Genet 56:1224–1233

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross mapping strategy and RAPD markers. Genetics 137:1121–1137

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

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135:205–211

Jansen RC (1996) A general Monte Carlo method for mapping multiple quantitative trait loci. Genetics 142:305–311

Jiang C, Zeng Z-B (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101:47–58

Knapp SJ, Holloway JL, Bridges WC (1995) Mapping dominant markers using F_2 matings. Theor Appl Genet 91:74–81

Knott SA, Neale DB, Sewell MM, Hally CS (1997) Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. Theor Appl Genet 94:810–820

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Martínez O, Curnow RN (1994) Missing markers when estimating quantitative trait loci using regression mapping. Heredity 73:198–206

Plomion C, Liu B-H, O'Malley DM (1996) Genetic analysis using trans-dominant linked markers in an F_2 family. Theor Appl Genet 93:1083–1089

Visscher PM, Thompson R, Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. Genetics 143 : 1013–1020
Williams JGK, Hanafey MK, Rafalski JA, Tingey SV (1993) Genetic analysis using random amplified polymorphic DNA markers. Methods Enzymol 218 : 704–740
Xie C, Gessler DDG, Xu S (1998) Sib mating designs for mapping quantitative trait loci. Genetica (in press)
Xu S (1995) A comment on the simple regression method for interval mapping. Genetics 141 : 1657–1659
Xu S (1996) Mapping quantitative trait loci using four-way crosses. Genet Res 68 : 175–181
Xu S (1998) Further investigation on the regression method of mapping quantitative trait loci. Heredity 80 : 364–373
Xu S, Gessler DDG (1998) Multipoint genetic mapping of quantitative trait loci using a variable number of sibs per family. Genet Res 71 : 73–83
Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136 : 1457–1468